บทที่ 6

สรุปผลและเสนอแนะ

สรุปผล

จากการวิจัยการออกแบบและพัฒนาส่วนจำเพาะการค้นข้อความไทยในเอกสารพีดีเอฟ พบว่า ปัญหาของการค้นข้อความไทยในเอกสารพีดีเอฟที่ไม่สามารถค้นข้อความไทยได้อย่างถูกต้อง เนื่องมาจาก เครื่องมือการค้นในโปรแกรมแสดงเอกสารพีดีเอฟไม่เข้าใจการเข้ารหัสที่แบบอักษรไทยใช้ และการเข้ารหัสตัวอักษรของแบบอักษรไทยในเอกสารพีดีเอฟเข้ารหัสไม่เป็นไปตามข้อกำหนดมาตรฐานการเข้ารหัสที่เอกสารพีดีเอฟกำหนด

เอกสารพีดีเอฟภาษาไทยที่เป็นปัญหาในการค้นข้อความ สามารถจำแนกออกได้เป็น 2 กลุ่มด้วยกัน คือ กลุ่มเอกสารพีดีเอฟที่แบบอักษรไทยถูกเข้ารหัสตัวอักษรใหม่หรือที่เรียกว่า เข้ารหัสโดยผู้ใช้กำหนดการเข้ารหัสตัวอักษรเอง แบบอักษรไทยในเอกสารพีดีเอฟที่ถูกเข้ารหัสตัวอักษรใหม่ สามารถจำแนกการเข้ารหัสตามลักษณะรหัสอักขระและชื่ออักขระ ได้ดังนี้

1. รหัสอักขระในแบบอักษรตรงตามมาตรฐาน มอก.620
2. รหัสอักขระในแบบอักษรลดลงจากมาตรฐาน มอก.620
3. รหัสอักขระในการเข้ารหัสไม่คงที่ ชื่ออักขระในการเข้ารหัสคงที่
4. รหัสอักขระและชื่ออักขระในแบบอักษรไม่คงที่

และกลุ่มเอกสารพีดีเอฟภาษาไทยที่ไม่มีข้อมูลการเข้ารหัสในพจนานุกรมแบบอักษร แบบอักษรไทยในเอกสารพีดีเอฟที่ระบุการเข้ารหัสแบบนี้ จะใช้การเข้ารหัสตัวอักษรในแฟ้มแบบอักษร

จากลักษณะและปัญหาของการเข้ารหัสแบบอักษรไทยในเอกสารพีดีเอฟ จึงออกแบบส่วนจำเพาะการค้นข้อความไทยในเอกสารพีดีเอฟ ให้มีกระบวนการทำงานดังนี้

1. วิเคราะห์การเข้ารหัสแบบอักษร

แบ่งการวิเคราะห์การเข้ารหัสแบบอักษรออกเป็น 2 ระดับ คือ

- วิเคราะห์การเข้ารหัสแบบอักษรโดยใช้ข้อมูลสภาพแวดล้อมที่มีอยู่ในเอกสาร เช่น ชื่อแบบอักษร ประเภทแบบอักษร ตารางความกว้างของอักขระ และเครื่องมือที่ใช้สร้างเอกสาร แต่ในบางเอกสารมีข้อมูลเหล่านี้ไม่เพียงพอในการวิเคราะห์การเข้ารหัสแบบอักษร จึงต้องทำการวิเคราะห์การเข้ารหัสในระดับถัดไป

- เมื่อเอกสารพีดีเอฟมีข้อมูลไม่เพียงพอในการวิเคราะห์การเข้ารหัสโดยใช้ข้อมูลสภาพแวดล้อมในเอกสารพีดีเอฟ จะทำการวิเคราะห์การเข้ารหัสแบบอักษรในเอกสารพีดีเอฟโดยการถามผู้ใช้ การวิเคราะห์แบบนี้จะนำข้อความในเอกสารมาแสดงข้อความ 2 ข้อความ ข้อความแรกจะใช้รหัสอักขระตามที่ใช้ในเอกสารพีดีเอฟ และข้อความที่ 2 จะบวกรหัสอักขระของข้อความแรกด้วยค่า 35_8

2. ถอดรหัสข้อความไทยในเอกสารพีดีเอฟ จะทำการคัดแยกข้อความออกมาจากเอกสารพีดีเอฟแล้วทำการแปลงรหัสอักขระให้ตรงตามข้อกำหนดมาตรฐาน มอก.620 โดยการถอดรหัสข้อความไทยในเอกสารพีดีเอฟจะมีอยู่ด้วยกัน 4 แบบ คือ

- ถอดรหัสข้อความไทยในเอกสารที่มีการเข้ารหัสตัวอักษรตรงตามมาตรฐาน มอก.620 วิธีนี้จะถอดรหัสข้อความ โดยทำการคัดแยกข้อความออกมาจากเอกสารพีดีเอฟ และใช้ค่ารหัสอักขระตามที่คัดแยกออกมาได้

- ถอดรหัสข้อความไทยในเอกสารที่มีการเข้ารหัสตัวอักษรลดลงจากมาตรฐาน มอก.620 วิธีนี้จะถอดรหัสข้อความ โดยทำการคัดแยกข้อความออกมาจากเอกสารพีดีเอฟ แล้วนำไปหารหัสตัวอักษรที่ต้องการในตารางที่เตรียมไว้

- ถอดรหัสข้อความไทยในเอกสารที่รหัสตัวอักษรที่ใช้ในการเข้ารหัสไม่คงที่แต่ชื่อตัวอักษรคงที่ วิธีนี้จะถอดรหัสข้อความโดยดูว่ารหัสตัวอักษรมีชื่อตัวอักษรว่าอย่างไร ถ้าชื่อตัวอักษรขึ้นต้นด้วย "afii" จะถอดรหัสตัวอักษรโดยใช้ตัวเลข 3 หลักด้านท้ายของชื่อตัวอักษร ลบด้วย 520 ถ้าชื่อตัวอักษร ขึ้นต้นด้วย "f7" จะถอดรหัสตัวอักษรโดยใช้ตัวเลข 2 หลักด้านท้าย ไปหารหัสตัวอักษรที่ต้องการในตารางที่เตรียมไว้

- ถอดรหัสข้อความไทยในเอกสารที่ไม่มีข้อมูลการเข้ารหัสในพจนานุกรมแบบอักษร วิธีนี้จะถอดรหัสข้อความ โดยทำการคัดแยกข้อความออกมาจากเอกสารพีดีเอฟเป็นรหัสยูนิโค้ด ค่ารหัสตัวอักษรที่ได้จะเป็นรหัสขนาด 2 ไบต์ ถ้ารหัสตัวอักษรไบต์แรกเท่ากับ "0E" จะถอดรหัสตัวอักษรโดยใช้รหัสตัวอักษรไบต์หลัง บวกด้วย 160 ถ้ารหัสตัวอักษรไบต์แรกเท่ากับ "f7" จะถอดรหัสตัวอักษรโดยนำค่ารหัสตัวอักษรไบต์หลัง ไปหารหัสตัวอักษรที่ต้องการในตารางที่เตรียมไว้

3. ค้นข้อความโดยวิธีการเปรียบเทียบสายอักขระ ทำการเปรียบอักขระข้อความที่ต้องการค้นกับอักขระที่ได้จากเอกสารพีดีเอฟโดยการเลื่อนอักขระไปด้านขวามือที่ละ 1 อักขระ แต่เนื่องจากข้อความที่ได้จากเอกสารพีดีเอฟอาจเป็นข้อความย่อยๆ การเปรียบเทียบข้อความจึงเป็นไปได้ 3 กรณี คือ

1) ไม่พบข้อความที่ต้องการค้นในข้อความที่ได้จากเอกสารพีดีเอฟ

2) พบข้อความที่ต้องการค้นในข้อความที่ได้จากเอกสารพีดีเอฟแต่พบเป็นข้อความย่อยของข้อความที่ต้องการค้น ในกรณีนี้จะนำข้อความย่อยที่ยังค้นไม่ครบมาค้นต่อในเอกสารพีดีเอฟ

3) พบข้อความที่ต้องการค้นในข้อความที่ได้จากเอกสารพีดีเอฟ

ส่วนจำเพาะค้นข้อความไทยในเอกสารพีดีเอฟ สามารถค้นข้อความได้ทั้งภาษาไทยและภาษาอังกฤษ ในเอกสารพีดีเอฟที่มีการเข้ารหัสแบบอักษรที่ไม่เป็นไปตามข้อกำหนดมาตรฐานที่เอกสารพีดีเอฟกำหนด ซึ่งเครื่องมือการค้นในโปรแกรมแสดงเอกสารพีดีเอฟโดยทั่วไปไม่สามารถจัดการได้ แต่ส่วนจำเพาะค้นข้อความไทยในเอกสารพีดีเอฟมีข้อจำกัด คือ

1. การวิเคราะห์การเข้ารหัสแบบอักษรในบางเอกสารต้องให้ผู้ใช้ช่วยวิเคราะห์

2. เอกสารพีดีเอฟที่มีการเข้ารหัสตัวอักษรไม่คงที่แต่ชื่อตัวอักษรที่ใช้ในการเข้ารหัสคงที่จะไม่สามารถค้นตัวอักษรได้ 2 ตัว คือ ตัวอักษรตัวที่ 13 และ ตัวที่ 32

3. ไม่สามารถแสดงสีทึบทับข้อความที่พบตามขนาดข้อความที่ค้นได้

ข้อเสนอแนะ

จากการทำวิทยานิพนธ์นี้พบว่า มีข้อเสนอแนะบางอย่างที่จะนำมาเป็นข้อเสนอเพื่อที่ผู้ที่สนใจการค้นข้อความไทยในเอกสารพีดีเอฟจะนำไปพัฒนาต่อให้ดียิ่งขึ้น คือ

1. พัฒนากระบวนการในการสร้างเอกสารพีดีเอฟ ให้เอกสารพีดีเอฟภาษาไทยใช้ข้อมูลการเข้ารหัสที่เป็นมาตราฐานตามที่เอกสารพีดีเอฟกำหนด เช่น การเข้ารหัสมาตราฐานวินโดวส์ เนื่องจาก การค้นข้อความในเอกสารพีดีเอฟขึ้นอยู่กับข้อมูลการเข้ารหัสของแบบอักษร แบบอักษรไทยในเอกสารพีดีเอฟที่พบเห็นในปัจจุบัน มีการเข้ารหัสที่ไม่เป็นไปตามมาตราฐานตามข้อกำหนดของเอกสารพีดีเอฟ ผู้ที่สนใจการค้นข้อความไทยในเอกสารพีดีเอฟควรจะพัฒนากระบวนการในขั้นตอนการสร้างเอกสารพีดีเอฟนี้ ให้ได้เอกสารพีดีเอฟภาษาไทยที่ใช้การเข้ารหัสแบบอักษรที่เป็นมาตราฐานตามที่เอกสารพีดีเอฟกำหนด จะทำให้การค้นข้อความไทยในเอกสารพีดีเอฟมีประสิทธิภาพดียิ่งขึ้น

2. ส่วนจำเพาะการค้นข้อความไทยในเอกสารพีดีเอฟที่ผู้วิจัยได้พัฒนาขึ้นนี้ สามารถนำไปใช้ในการค้นข้อความไทยในเอกสารพีดีเอฟได้ครั้งละหนึ่งเอกสาร ผู้ที่สนใจการค้นข้อความไทยในเอกสารพีดีเอฟ สามารถพัฒนาเพิ่มเติมในเรื่องการค้นข้อความไทยในกลุ่มเอกสารพีดีเอฟที่ต้องการได้