

การออกแบบและพัฒนาส่วนจำเพาะการค้นข้อความไทยในเอกสารพีดีเอฟ

นาย สุรพงษ์ เชาวน์เชี่ยวชาญ



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2543

ISBN 974-346-959-1

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

I 19970 274

A DESIGN AND DEVELOPMENT OF A THAI TEXT SEARCH MODULE IN PDF FILES

Mr. Surapong Chaocheawchan

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2000

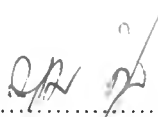
ISBN 974-346-959-1

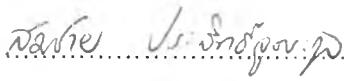
หัวข้อวิทยานิพนธ์ : การออกแบบและพัฒนาส่วนจำเพาะการค้นข้อความไทยในเอกสาร
พีดีเอฟ
โดย : นายสุรพงษ์ เชาว์เชี่ยวชาญ
ภาควิชา : วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร.สมชาย ประสิทธิ์จตุระกุล


คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้หัวข้อวิทยานิพนธ์ฉบับนี้เป็นส่วน
หนึ่งของการศึกษาตามหลักสูตรปริญญาโทบัณฑิต

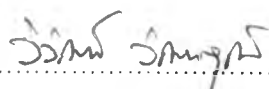

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สมศักดิ์ ปัญญาแก้ว)

คณะกรรมการสอบวิทยานิพนธ์


..... ประธานกรรมการ
(อาจารย์ ดร.บุญเสริม กิจศิริกุล)


..... อาจารย์ที่ปรึกษา
(ผู้ช่วยศาสตราจารย์ ดร.สมชาย ประสิทธิ์จตุระกุล)


..... กรรมการ
(ผู้ช่วยศาสตราจารย์ นงลักษณ์ โควาวีสารัช)


..... กรรมการ
(ผู้ช่วยศาสตราจารย์ วิวัฒน์ วัฒนาวุฒิ)

สุรพงษ์ เชาวน์เขียวชาญ : การออกแบบและพัฒนาส่วนจำเพาะการค้นหาข้อความไทยในเอกสารพีดีเอฟ (A DESIGN AND DEVELOPMENT OF A THAI TEXT SEARCH MODULE IN PDF FILES) อ. ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จูตระกูล, 97 หน้า ISBN 974-346-959-1.

วิทยานิพนธ์นี้นำเสนอการออกแบบและพัฒนาส่วนจำเพาะซึ่งใช้สำหรับการค้นหาข้อความไทยในเอกสารพีดีเอฟ ส่วนจำเพาะนี้มีหน้าที่หลักคือการถอดรหัสอักขระ การเปรียบเทียบสายอักขระ และการแสดงตำแหน่งในเอกสารที่ค้นพบ ความซับซ้อนของส่วนจำเพาะนี้อยู่ที่ขั้นตอนการถอดรหัส ทั้งนี้เนื่องจากเครื่องมือการสร้างเอกสารพีดีเอฟในปัจจุบันไม่สนับสนุนการเข้ารหัสภาษาไทยที่เป็นมาตรฐาน ดังนั้นอักขระไทยต่างๆในเอกสารพีดีเอฟจึงถูกเข้ารหัสในหลากหลายรูปแบบ ขั้นตอนการถอดรหัสอาศัยข้อมูลของแบบอักษรชื่ออักขระและสภาพแวดล้อมที่สร้างเอกสารพีดีเอฟนั้นๆ ประกอบการวิเคราะห์การถอดรหัส การพัฒนาอาศัยชุดพัฒนาส่วนจำเพาะที่ใช้ได้กับซอฟต์แวร์แสดงเอกสารพีดีเอฟอะโครเบต ส่วนจำเพาะสามารถค้นหาข้อความไทยในเอกสารพีดีเอฟที่มีการเข้ารหัสภาษาไทยในรูปแบบคงที่ได้ทุกรูปแบบ

ภาควิชา วิศวกรรมคอมพิวเตอร์
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์
ปีการศึกษา ... 2543

ลายมือชื่อนิสิต สุรพงษ์ เชาวน์เขียวชาญ
ลายมือชื่ออาจารย์ที่ปรึกษา สมชาย ประสิทธิ์จูตระกูล
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

4070477321: MAJOR COMPUTER SCIENCE

KEY WORD: FONT ENCODING / STRING PATTERN MATCHING / TEXT EXTRACTION /
ELECTRONIC DOCUMENT

SURAPONG CHAOCHAEWCHAN : A DESIGN AND DEVELOPMENT OF A
THAI TEXT SEARCH MODULE IN PDF FILES. THESIS ADVISOR :

ASSISTANT PROFESSOR SOMCHAI PRASITJUTRAKUL. 97 pp. ISBN 974-
346-959-1

This thesis presents a design and development of a Thai text search module in PDF files. The Objectives of this module are to decode characters, match strings, and highlight the matched strings. The complexity of the module is in the decoding step since current PDF creation tools do not support standard Thai character encoding. As a result, Thai characters are encoded in many different formats. The decoding step uses font description, character names, and information related to tools and environment used for generating PDF files for analyzing encoded characters. The module was developed using Acrobat PDF software development kit. The module can search Thai text in any fixed-format encoded PDF files.

DepartmentComputer Engineering.....	Student's signature <i>Surapong Chaochewchan</i>
Field of study Computer Science.....	Advisor's signature <i>Somchai Prasitjutrakul</i>
Academic year ...2543.....	Co-advisor's signature



กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลือและคำแนะนำอย่างดียิ่งของผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จตุระกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งได้ให้คำแนะนำและข้อคิดเห็นที่เป็นประโยชน์ในการวิจัย นอกจากนี้ยังมีคณะกรรมการที่ได้ช่วยตรวจสอบและให้คำแนะนำที่เป็นประโยชน์ต่อการแก้ไขปรับปรุงเพื่อให้วิทยานิพนธ์นี้ถูกต้องสมบูรณ์ยิ่งขึ้น ผู้วิจัยจึงขอขอบ คุณมา ณ ที่นี้

นอกจากนี้ ผู้วิจัยขอขอบคุณ คุณวรากร พุศิริพงษ์ ที่ให้การสนับสนุนอุปกรณ์เครื่องพิมพ์ พร้อมทั้ง คุณวิภา พิณรัตน์ และคุณสุรัชย์ เชาว์เชี่ยวชาญ ที่ให้การสนับสนุนทางด้านอุปกรณ์คอมพิวเตอร์และทุนทรัพย์ที่ใช้ในการวิจัย รวมทั้งขอกราบขอบพระคุณ บิดา-มารดา ที่ให้กำลังใจแก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา

สารบัญ

	หน้า
บทคัดย่อภาษาไทย.....	ง
บทคัดย่อภาษาอังกฤษ.....	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญ.....	ช
สารบัญตาราง.....	ฌ
สารบัญรูป.....	ญ
บทที่	
1. บทนำ.....	1
1.1 ความเป็นมาและความสำคัญของปัญหา.....	1
1.2 วัตถุประสงค์.....	4
1.3 ขอบเขตการวิจัย.....	4
1.4 ขั้นตอนการวิจัย.....	5
1.5 ผลที่คาดว่าจะได้รับ.....	5
2. แนวคิดและทฤษฎี.....	6
2.1 เอกสารพีดีเอฟ.....	6
2.1.1 การสร้างเอกสารพีดีเอฟ.....	6
2.1.2 โครงสร้างแฟ้มเอกสารพีดีเอฟ.....	7
2.1.3 แบบอักษรในเอกสารพีดีเอฟ.....	11
2.1.4 การเข้ารหัสตัวอักษร.....	20
2.2 มาตรฐานการเข้ารหัสอักขระไทย.....	22
2.2.1 มาตรฐาน มอก.620.....	22
2.2.2 มาตรฐานแบบอักษรไทยบนระบบปฏิบัติการวินโดวส์.....	24
2.3 การค้นข้อความโดยวิธีการเปรียบเทียบสายอักขระ.....	26
2.3.1 การเปรียบเทียบสายอักขระตามแนวคิดของ Brute-Force.....	26
2.3.2 การเปรียบเทียบสายอักขระตามแนวคิดของ Knuth-Morris-Pratt.....	27
2.3.3 การเปรียบเทียบสายอักขระตามแนวคิดของ Boyer และ Moore.....	27

3. การเข้าและถอดรหัสข้อความไทยในเอกสารพีดีเอฟ.....	28
3.1 การเข้ารหัสตัวอักษรของแบบอักษรไทยในเอกสารพีดีเอฟ.....	28
3.1.1 การเข้ารหัสของแบบอักษรไทยแบบผู้กำหนดการเข้ารหัสตัวอักษรเอง	28
3.1.2 การเข้ารหัสแบบอักษรไทยที่ใช้ข้อกำหนดการเข้ารหัสในแฟ้มแบบอักษร	32
3.2 การถอดรหัสข้อความไทยในเอกสารพีดีเอฟ	33
3.2.1 การวิเคราะห์การเข้ารหัสแบบอักษรไทยในเอกสารพีดีเอฟ.....	34
3.2.2 การคัดแยกข้อความไทยแล้วแปลงรหัสอักขระให้ตรงตามข้อกำหนด มอก.620	39
4. กระบวนการค้นข้อความไทยในเอกสารพีดีเอฟ	43
4.1 การค้นข้อความไทยในเอกสารพีดีเอฟโดยวิธีการเปรียบเทียบสายอักขระ	43
4.2 การแสดงตำแหน่งข้อความที่ค้นพบในเอกสารพีดีเอฟ	47
5. การพัฒนาส่วนจำเพาะการค้นข้อความไทยในเอกสารพีดีเอฟ.....	48
5.1 ความสัมพันธ์ของคลาสในส่วนจำเพาะการค้นข้อความไทยในเอกสารพีดีเอฟ.....	48
5.2 การพัฒนาคลาสและฟังก์ชันทำงานในส่วนจำเพาะการค้นข้อความไทยในเอกสารพีดีเอฟ	50
5.2.1 คลาส CPDFPlugins.....	50
5.2.2 คลาส CAboutDlg.....	52
5.2.3 คลาส CFindDlg.....	52
5.2.4 คลาส CThaiPDF	54
5.2.5 คลาส CProgressDlg	58
5.2.6 คลาส CFontRegDlg	59
6. สรุปผลและเสนอแนะ	62
รายการอ้างอิง.....	65
ภาคผนวก.....	66
ภาคผนวก ก. มาตรฐานการเข้ารหัสแบบอักษรในเอกสารพีดีเอฟ	67
ภาคผนวก ข. โปรแกรมส่วนจำเพาะการค้นข้อความไทยในเอกสารพีดีเอฟ.....	72
ประวัติผู้เขียน.....	97

สารบัญตาราง

		หน้า
ตารางที่ 1	ข้อมูลส่วนต่างๆของแบบอักษรประเภทที่ 1.....	12
ตารางที่ 2	ข้อมูลส่วนต่างๆของแบบอักษรส่วนขยายประเภทที่ 1	14
ตารางที่ 3	ข้อมูลส่วนต่างๆของแบบอักษรประเภทที่ 3	16
ตารางที่ 4	ข้อมูลส่วนต่างๆของแบบอักษรประเภททรูไทย.....	18
ตารางที่ 5	ตารางการเข้ารหัสแบบอักษรไทยในระบบปฏิบัติการวินโดวส์.....	24
ตารางที่ 6	ตารางการเข้ารหัสแบบอักษรไทยในเอกสารพีดีเอฟ ที่รหัสตัวอักษรตรงตามมาตรฐาน มอก.620	29
ตารางที่ 7	ตารางการเข้ารหัสแบบอักษรไทยในเอกสารพีดีเอฟ ที่รหัสตัวอักษรลดลงมาตรฐาน มอก.620.....	30
ตารางที่ 8	ตารางชื่อตัวอักษรที่แบบอักษรไทยใช้ในเอกสารพีดีเอฟ.....	31
ตารางที่ 9	ตารางรหัสยูนิโคดที่แบบอักษรไทยใช้ในเอกสารพีดีเอฟ.....	33
ตารางที่ 10	อักขระในภาษาไทยที่มีรหัสอักขระมากกว่า 1 รหัสอักขระ	35
ตารางที่ 11	ตารางถอดรหัสข้อความไทยในเอกสารพีดีเอฟที่มี การเข้ารหัสตัวอักษรลดลง จากมาตรฐาน มอก.620	39
ตารางที่ 12	ตารางถอดรหัสข้อความไทยในเอกสารพีดีเอฟที่ แบบอักษรเข้ารหัสด้วยชื่อตัวอักษร "f7".....	40

สารบัญรูป

		หน้า
รูปที่	1	เอกสารพีดีเอฟศูนย์คอมพิวเตอร์คณะวิศวกรรมศาสตร์ จุฬาฯ..... 2
รูปที่	2	แนวคิดกระบวนการค้นข้อความไทยในเอกสารพีดีเอฟ..... 3
รูปที่	3	กระบวนการสร้างเอกสารพีดีเอฟโดยวิธีพีดีเอฟไรเตอร์..... 6
รูปที่	4	กระบวนการสร้างเอกสารพีดีเอฟโดยวิธีดีสทิวเลอร์..... 7
รูปที่	5	โครงสร้างแฟ้มเอกสารพีดีเอฟ..... 7
รูปที่	6	ตัวอย่างโครงสร้างแฟ้มเอกสารพีดีเอฟ..... 8
รูปที่	7	โครงสร้างต้นไม้ของวัตถุที่ใช้ในการแสดงเอกสารพีดีเอฟ..... 9
รูปที่	8	ตัวอย่างตารางอ้างอิงในแฟ้มเอกสารพีดีเอฟ..... 10
รูปที่	9	ตัวอย่างข้อมูลส่วนท้ายของแฟ้มเอกสารพีดีเอฟ..... 10
รูปที่	10	ตัวอย่างแบบอักษรประเภทที่ 1 ในเอกสารพีดีเอฟ..... 13
รูปที่	11	ตัวอย่างแบบอักษรส่วนขยายประเภทที่ 1 ในเอกสารพีดีเอฟ..... 15
รูปที่	12	ตัวอย่างชุดแบบตัวประเภทที่ 3 ในเอกสารพีดีเอฟ..... 17
รูปที่	13	ตัวอย่างชุดแบบตัวประเภททรูไทป์ ในเอกสารพีดีเอฟ..... 19
รูปที่	14	วิธีการแสดงตัวอักษรในเอกสารพีดีเอฟ โดยใช้ความสัมพันธ์ระหว่างรหัสตัวอักษรและชื่อตัวอักษร..... 20
รูปที่	15	ตัวอย่างการเข้ารหัสตัวอักษรในเอกสารพีดีเอฟ ตามข้อกำหนดมาตรฐานวินโดวส์..... 21
รูปที่	16	ตัวอย่างการเข้ารหัสตัวอักษรในเอกสารพีดีเอฟ แบบผู้ใ้กำหนดการเข้ารหัสตัวอักษรเอง..... 21
รูปที่	17	ตัวอย่างการเข้ารหัสตัวอักษรในเอกสารพีดีเอฟ แบบใช้ข้อกำหนดที่มีอยู่ในแฟ้มแบบอักษร..... 22
รูปที่	18	ข้อกำหนดการเข้ารหัสตัวอักษรภาษาไทยตามมาตรฐาน มอก.620..... 23
รูปที่	19	ผังงานการวิเคราะห์การเข้ารหัสแบบอักษรไทย ในเอกสารพีดีเอฟ..... 38
รูปที่	20	กระบวนการคัดแยกข้อความออกมาจากเอกสารพีดีเอฟ..... 42
รูปที่	21	ผังงานกระบวนการเปรียบเทียบข้อความในส่วนจำเพาะ ค้นข้อความไทย..... 46

รูปที่	22	ผังความสัมพันธ์ของคลาสในส่วนจำเพาะการค้นข้อความไทย _____	49
รูปที่	23	หน้าต่างของคลาส CAboutDlg.....	52
รูปที่	24	หน้าต่างของคลาส CFindDlg _____	52
รูปที่	25	หน้าต่างของคลาส CProgressDlg	58
รูปที่	26	หน้าต่างของคลาส CFontRegDlg	60