การออกแบบและพัฒนาส่วนจำเพาะการค้นข้อความไทยในเอกสารพีดีเอฟ

นาย สุรพงษ์ เชาว์เชี่ยวชาญ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2543

ISBN 974-346-959-1

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	: การออกแบบและพัฒนาส่วนจำเพาะการค้นข้อความไทยในเอกสารพีดีเอฟ
โดย	: นายสุรพงษ์ เชาว์เชี่ยวชาญ
ภาควิชา	: วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	: ผู้ช่วยศาสตราจารย์ ดร.สมชาย ประสิทธิ์จูตระกูล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

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

คณะกรรมการสอบวิทยานิพนธ์

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

..... อาจารย์ที่ปรึกษา
(ผู้ช่วยศาสตราจารย์ ดร.สมชาย ประสิทธิ์จูตระกูล)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ นงลักษณ์ โควาวิสารัช)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ วิวัฒน์ วัฒนาวุฒิ)

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

วิทยานิพนธ์นี้นำเสนอการออกแบบและพัฒนาส่วนจำเพาะซึ่งใช้สำหรับการค้นข้อความไทยในเอกสารพีดีเอฟ ส่วนจำเพาะนี้มีหน้าที่หลักคือการถอดรหัสอักขระ การเปรียบเทียบสายอักขระ และการแสดงตำแหน่งในเอกสารที่ค้นพบ ความซับซ้อนของส่วนจำเพาะนี้อยู่ที่ขั้นตอนการถอดรหัส ทั้งนี้เนื่องจากเครื่องมือการสร้างเอกสารพีดีเอฟในปัจจุบันไม่สนับสนุนการเข้ารหัสภาษาไทยที่เป็นมาตรฐาน ดังนั้นอักขระไทยต่างๆในเอกสารพีดีเอฟจึงถูกเข้ารหัสในหลากหลายรูปแบบ ขั้นตอนการถอดรหัสอาศัยข้อมูลของแบบอักษรชื่ออักขระและสภาพแวดล้อมที่สร้างเอกสารพีดีเอฟนั้นๆ ประกอบการวิเคราะห์การถอดรหัส การพัฒนาอาศัยชุดพัฒนาส่วนจำเพาะที่ใช้ได้กับซอฟต์แวร์แสดงเอกสารพีดีเอฟอะโครแบต ส่วนจำเพาะสามารถค้นข้อความไทยในเอกสารพีดีเอฟที่มีการเข้ารหัสภาษาไทยในรูปแบบคงที่ได้ทุกรูปแบบ

ภาควิชา ....วิศวกรรมคอมพิวเตอร์.....
สาขาวิชา ....วิทยาศาสตร์คอมพิวเตอร์.....
ปีการศึกษา ...2543.....

ลายมือชื่อนิสิต ...สุรพงษ์ เชาว์เชี่ยวชาญ.....
ลายมือชื่ออาจารย์ที่ปรึกษา ...สมชาย ประสิทธิ์จูตระกูล...
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม .....

4070477321: MAJOR COMPUTER SCIENCE
KEY WORD: FONT ENCODING / STRING PATTERN MATCHING / TEXT EXTRACTION / ELECTRONIC DOCUMENT
SURAPONG CHAOCHEAWCHAN : 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.

Department .....Computer Engineering.....	Student's signature Surapong Chaocheawchan
Field of study .... Computer Science.....	Advisor's signature .....Somchai Prasitjutrakul
Academic year ...2543.....	Co-advisor's signature .....

กิตติกรรมประกาศ

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

นอกจากนี้ ผู้วิจัยขอขอบคุณ คุณวรากร ฟูศิริพงษ์ ที่ให้การสนับสนุนอุปกรณ์เครื่องพิมพ์ พร้อมทั้ง คุณวิภา พิณรัตน์ และคุณสุรชัย เชาว์เชี่ยวชาญ ที่ให้การสนับสนุนทางด้านอุปกรณ์คอมพิวเตอร์และทุนทรัพย์ที่ใช้ในการวิจัย รวมทั้งขอกราบขอบพระคุณ บิดา-มารดา ที่ให้กำลังใจแก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา

สารบัญ

สารบัญตาราง

สารบัญรูป