ระบบการค้นคืนข้อความภาษาไทยโดยใช้แฟ้มข้อมูลผกผัน

นาย วิฑูรย์ กัลยาณวัฒน์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2540

ISBN 974-637-663-2

THAI TEXT RETRIEVAL SYSTEM USING INVERTED FILES

Mr. Witoon Kanlayanawat

A Thesis Submitted in Partial FulFillment of the Requirements
For the Degree of Master of Engineering in Computer Engineering
Department of Computer Engineering
Graduate School
Chulalongkorn University
Academic Year 1997

974-637-663-2

หัวข้อวิทยานิพนธ์	ระบบการค้นคืนข้อความภาษาไทยโดยใช้แฟ้มข้อมูลผกผัน
โดย	นาย วิฑูรย์ กัลยาณวัฒน์
ภาควิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จูตระกูล

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ นายแพทย์ ศุภวัฒน์ ชุติวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร. วันชัย ริ้วไพบูลย์)

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

..... กรรมการ
(อาจารย์ ดร. บุญเสริม กิจศิริกุล)

..... กรรมการ
(รองศาสตราจารย์ ดร. วิทยา วัชระวิทยากุล)

วิฑูรย์ กัลยาณวัฒน์ : ระบบการค้นคืนข้อความภาษาไทยโดยใช้แฟ้มข้อมูลผกผัน (THAI TEXT RETRIEVAL SYSTEM USING INVERTED FILES) อ.ที่ปรึกษา : ผศ. ดร. สมชาย ประสิทธิ์จูตระกูล , 68 หน้า. ISBN 974-637-663-2.

วิทยานิพนธ์ฉบับนี้นำเสนอขั้นตอนวิธีการจัดทำดัชนีสำหรับระบบสืบค้นข้อความไทยที่ใช้โครงสร้างแฟ้มข้อมูลแบบผกผัน โดยที่เอกสารต่างๆ ที่ได้รับมานั้นสามารถมีคำที่ไม่มีอยู่ในพจนานุกรมของระบบได้ ปัญหานี้เกิดขึ้นจากการเขียนข้อความในภาษาไทยที่ไม่มีตัวกำหนดขอบเขตระหว่างคำ โดยอาศัยพจนานุกรมของระบบ ขั้นตอนวิธีที่นำเสนอนี้หาคำที่ยาวสุดต่างๆ ที่มีในพจนานุกรมที่ปรากฏในข้อความ จากนั้นสร้างกราฟที่แทนการติดกันและการทับกันของคำต่างๆ ในข้อความ โดยที่เส้นทางที่สั้นสุดในกราฟนี้แทนกลุ่มที่เล็กสุดของคำในข้อความที่เมื่อเลือกแล้วจะลดจำนวนลายอักษรย่อยที่ไม่รู้จักให้ปรากฏขึ้นเป็นจำนวนน้อยที่สุด ลายอักษรย่อยเหล่านี้จะถูกเทียบกับพยางค์ต่างๆ ในข้อความ โดยการใช้ขั้นตอนวิธีการแบ่งพยางค์แบบใช้กฎ ค่าต่างๆ ที่ได้บนเส้นทางสั้นสุดของกราฟ และพยางค์ต่างๆ ที่ได้จากการเทียบกับลายอักษรย่อยที่ไม่เป็นคำที่รู้จัก จะเป็นกลุ่มของคำสำคัญในการจัดทำดัชนีของข้อความที่ได้รับ ผลการทดลองแสดงให้เห็นว่าจำนวนคำสำคัญที่หาได้นั้นลดจากจำนวนคำทั้งหมดที่หาได้จากข้อความประมาณ 72 %

ภาควิชา วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต วิฑูรย์ กัลยาณวัฒน์
สาขาวิชา วิศวกรรมคอมพิวเตอร์	ลายมือชื่ออาจารย์ที่ปรึกษา สมชาย ประสิทธิ์จูตระกูล
ปีการศึกษา 2540	ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

CS18831 : MAJOR COMPUTER ENGINEERING
KEY WORD: INVERTED FILES / RETRIEVAL / THAI
WITOON KANLAYANAWAT : THAI TEXT RETRIEVAL SYSTEM USING INVERTED FILES. THESIS ADVISOR : ASSIST. PROF. SOMCHAI PRASITJUTRAKUL, Ph.D 68 pp. ISBN 974-637-663-2.

This thesis presents an automatic indexing algorithm for inverted-file-based Thai text retrieval system where given documents can have words that are unknown to the system's dictionary. The problem arises from the fact that there is no explicit inter-word delimiter in Thai text. By using system dictionary, the algorithm first finds a set of recognizable words that maximally match all the semi-infinite substrings of a given text. It then constructs an adjacent-overlapping graph whose a shortest path represents a smallest list of known words minimizing unknown substrings of the text. The unknown substrings are matched with the set of syllables obtained from a rule-based syllable segmentation of the text. The words on the shortest path of the adjacent-overlapping graph and the matched syllables are then used as keywords for indexing of the given text. Experimental results showed that the number of keywords obtained is approximately 72% less compared to the number obtained by using matching-all-known-words technique.

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต วิทูรย์ กัลยาณวัฒน์

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา สมชาย ประสิทธิ์จุตระกูล

ปีการศึกษา 2540 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของผู้ช่วยศาสตราจารย์ สมชาย ประสิทธิ์จูตระกูล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้คำแนะนำและข้อคิดเห็นต่าง ๆ ในการวิจัยมาด้วยดีตลอด และเนื่องจากการวิจัยครั้งนี้ได้รับทุนงบประมาณแผ่นดิน คณะวิศวกรรมศาสตร์ ปี พ.ศ. 2540 จึงขอขอบคุณมา ณ ที่นี้ด้วย

ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา-มารดา ซึ่งสนับสนุนในด้านการเงินและให้กำลังใจแก่ผู้วิจัยเสมอจนสำเร็จการศึกษา

สารบัญ

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า