

ระบบการค้นคืนข้อความภาษาไทยโดยใช้แฟ้มข้อมูลผกผัน



นาย วิฑูรย์ กัลยานวัฒนา

สถาบันวิทยบริการ

จุฬาลงกรณ์มหาวิทยาลัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2540

ISBN 974-637-663-2

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

10 ต.ค. 2544

I17922677

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

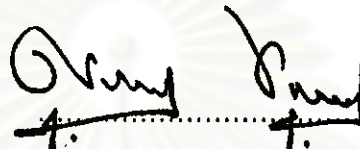
Academic Year 1997

974-637-663-2

หัวข้อวิทยานิพนธ์
โดย
ภาควิชา
อาจารย์ที่ปรึกษา

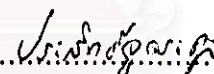
ระบบการค้นคืนข้อความภาษาไทยโดยใช้แฟ้มข้อมูลผกผัน
นาย วิฑูรย์ กัลยาณวัฒน์
วิศวกรรมคอมพิวเตอร์
ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จูตระกูล

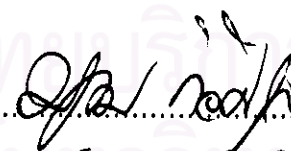
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้บัณฑิตวิทยาลัยรับนี้เป็นส่วน
หนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต

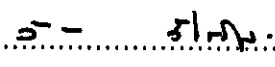

..... คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ นายแพทย์ สุภวัฒน์ ชุตินวงศ์)

คณะกรรมการสอบวิทยานิพนธ์


..... ประธานกรรมการ
(รองศาสตราจารย์ ดร. วันชัย ธีวไพบูลย์)

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


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

.....  กรรมการ
(รองศาสตราจารย์ ดร. วิทยา วัชรวิทยากุล)

วิทยุर्थ กัลยาณวิวัฒน์ : ระบบการค้นหาข้อความภาษาไทยโดยใช้แฟ้มข้อมูลผกผัน (THAI TEXT RETRIEVAL SYSTEM USING INVERTED FILES) อ.ที่ปรึกษา : ผศ. ดร. สมชาย ประสิทธิ์จตุระกุล , 68 หน้า. ISBN 974-637-663-2.

วิทยานิพนธ์ฉบับนี้นำเสนอขั้นตอนวิธีการจัดทำดัชนีสำหรับระบบสืบค้นข้อความไทยที่ใช้โครงสร้างแฟ้มข้อมูลแบบผกผัน โดยที่เอกสารต่างๆ ที่ได้รับมานั้นสามารถมีคำที่ไม่มีอยู่ในพจนานุกรมของระบบได้ ปัญหานี้เกิดขึ้นจากการเขียนข้อความในภาษาไทยที่ไม่มีตัวกำหนดขอบเขตระหว่างคำ โดยอาศัยพจนานุกรมของระบบ ขั้นตอนวิธีที่นำเสนอนี้หาคำที่ยาวสุดต่างๆ ที่มีในพจนานุกรมที่ปรากฏในข้อความ จากนั้นสร้างกราฟที่แทนการติดกันและการทับกันของคำต่างๆ ในข้อความ โดยที่เส้นทางที่สั้นสุดในกราฟนี้แทนกลุ่มที่เล็กสุดของคำในข้อความที่เมื่อเลือกแล้วจะลดจำนวนสายอักขระย่อยที่ไม่รู้จักให้ปรากฏขึ้นเป็นจำนวนน้อยที่สุด สายอักขระย่อยเหล่านี้จะถูกเทียบกับพยางค์ต่างๆ ในข้อความ โดยการใช้ขั้นตอนวิธีการแบ่งพยางค์แบบใช้กฎ คำต่างๆ ที่ได้บนเส้นทางสั้นสุดของกราฟ และพยางค์ต่างๆ ที่ได้จากการเทียบ กับสายอักขระย่อยที่ไม่เป็นคำที่รู้จัก จะเป็นกลุ่มของคำสำคัญในการจัดทำดัชนีของข้อความที่ได้รับ ผลการทดลองแสดงให้เห็นว่าจำนวนคำสำคัญที่หาได้นั้นลดลงจากจำนวนคำทั้งหมดที่หาได้จากข้อความประมาณ 72 %

สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา วิศวกรรมคอมพิวเตอร์
สาขาวิชา วิศวกรรมคอมพิวเตอร์
ปีการศึกษา 2549

ลายมือชื่อนิติบัตร วิทยุर्थ กัลยาณวิวัฒน์
ลายมือชื่ออาจารย์ที่ปรึกษา อดิสรณ์ ประสิทธิ์จตุระกุล
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

C818831 : 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.



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา.....วิศวกรรมคอมพิวเตอร์.....

สาขาวิชา.....วิศวกรรมคอมพิวเตอร์.....

ปีการศึกษา..... 2540.....

ลายมือชื่อนิสิต..... กฤษณ์ ภัฆานวัฒน์.....

ลายมือชื่ออาจารย์ที่ปรึกษา..... สอนชัย ประสิทธิ์จตุรกุล.....

ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของผู้ช่วย
ศาสตราจารย์ สมชาย ประสิทธิ์จตุระกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้คำแนะนำ
และข้อคิดเห็นต่าง ๆ ในการวิจัยมาด้วยดีตลอด และเนื่องจากการวิจัยครั้งนี้ได้รับทุนงบประมาณ
แผ่นดิน คณะวิศวกรรมศาสตร์ ปี พ.ศ. 2540 จึงขอขอบคุณมา ณ ที่นี้ด้วย
ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา-มารดา ซึ่งสนับสนุนในด้านการเงินและให้
กำลังใจแก่ผู้วิจัยเสมอจนสำเร็จการศึกษา



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญ

| | หน้า |
|---|------|
| บทคัดย่อภาษาไทย..... | ง |
| บทคัดย่อภาษาอังกฤษ..... | จ |
| กิตติกรรมประกาศ..... | ฉ |
| สารบัญ..... | ช |
| สารบัญตาราง..... | ญ |
| สารบัญภาพ..... | ฎ |
| บทที่ | |
| 1. บทนำ..... | 1 |
| ความเป็นมาและปัญหา..... | 1 |
| วัตถุประสงค์ของวิทยานิพนธ์..... | 2 |
| ขอบเขตและเงื่อนไขของวิทยานิพนธ์..... | 2 |
| ขั้นตอนการวิจัย..... | 2 |
| ประโยชน์ที่คาดว่าจะได้รับ..... | 3 |
| โครงสร้างของวิทยานิพนธ์..... | 3 |
| 2. ภาษาไทยกับระบบสืบค้นข้อมูลโดยใช้แฟ้มข้อมูลผกผัน..... | 5 |
| การแบ่งพยางค์โดยใช้กฎ..... | 5 |
| การแบ่งคำด้วยพจนานุกรม..... | 5 |
| การใช้พจนานุกรมในการดึงคำภาษาไทย..... | 6 |
| กรณีที่ 1..... | 7 |
| กรณีที่ 2..... | 9 |
| ลักษณะอัลกอริทึมการจัดทำดัชนีที่นำมาใช้กับระบบสืบค้นข้อมูล ที่ใช้แฟ้มข้อมูลผกผัน..... | 17 |
| 3. แนวคิดและการออกแบบอัลกอริทึมการจัดทำดัชนีภาษาไทย..... | 18 |
| แนวความคิดที่ 1..... | 18 |

| | |
|---|----|
| แนวความคิดที่ 2..... | 20 |
| แนวความคิดที่ 3..... | 22 |
| แนวความคิดที่ 4..... | 24 |
| แนวความคิดที่ 5..... | 26 |
| อัลกอริทึมการจัดทำดัชนีภาษาไทย | 26 |
| ชั้นตอนที่ 1..... | 26 |
| ชั้นตอนที่ 2..... | 27 |
| ชั้นตอนที่ 3..... | 30 |
| ชั้นตอนที่ 4..... | 31 |
| อัลกอริทึมการจัดทำดัชนีภาษาไทยโดยใช้ sistring กับคำที่ไม่มีในพจนานุกรม..... | 35 |
| ชั้นตอนที่ 1..... | 35 |
| ชั้นตอนที่ 2..... | 35 |
| ชั้นตอนที่ 3..... | 36 |
| ชั้นตอนที่ 4..... | 36 |
| 4. การวิเคราะห์ผลการทำงาน..... | 39 |
| จำนวนคำหลักในแต่ละชั้นตอนที่ได้จากอัลกอริทึมการจัดทำดัชนี | 40 |
| การเปรียบเทียบจำนวนของคำนำหน้าที่ได้ในชั้นตอนที่ 2 ของอัลกอริทึมการจัดทำดัชนี | 42 |
| การเปรียบเทียบจำนวนคำหลักที่ได้จากอัลกอริทึมต่างๆ | 43 |
| ประสิทธิภาพในแต่ละชั้นตอนของอัลกอริทึม | 45 |
| 5. บทสรุปและข้อเสนอแนะ | 47 |
| สรุปการทำงานของอัลกอริทึมการจัดทำดัชนี..... | 47 |
| ข้อเสนอแนะ | 49 |
| รายการอ้างอิง | 51 |
| ภาคผนวก ก. การแบ่งคำภาษาไทยในโปรแกรมซียูโรท์เตอร์ | 55 |
| ลักษณะของตัวอักษรภาษาไทย..... | 56 |
| โครงสร้างโดยทั่วไปของคำในภาษาไทย | 57 |
| การสร้างกฎเกณฑ์ในการแบ่งคำ..... | 58 |
| ภาคผนวก ข..... | 61 |

| | |
|--|----|
| ตัวอย่างข้อมูลประเภทโคลงกลอน..... | 62 |
| ตัวอย่างข้อมูลประเภทข่าว..... | 62 |
| ตัวอย่างข้อมูลประเภทเนื้อเพลง | 63 |
| ตัวอย่างข้อมูลประเภทข้อสอบเข้ามหาวิทยาลัย..... | 64 |
| ภาคผนวก ค..... | 65 |
| Sistring..... | 66 |
| Tries..... | 66 |
| ประวัติผู้เขียนวิทยานิพนธ์ | 68 |



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญตาราง

| | หน้า |
|---|------|
| ตารางที่ 3.1 ตัวอย่างของ w_i ที่ได้จาก sistring..... | 27 |
| ตารางที่ 3.2 คำน้่านักของการทับกัน และการต่อกันพอดีของคำ..... | 28 |
| ตารางที่ 3.3 คำน้่านัก..... | 29 |
| ตารางที่ 3.4 ตัวอย่างของ x_i ที่ได้จาก sistring..... | 32 |
| ตารางที่ 4.1 ขนาดของข้อมูลที่ใช้ในการทดสอบ | 39 |
| ตารางที่ 4.2 ผลที่ได้จากการทดลอง แสดงเป็นเปอร์เซ็นต์ โดยให้ขั้นตอนแรกเท่ากับ 100 เปอร์เซ็นต์..... | 40 |
| ตารางที่ 4.3 ข้อมูลที่ได้จากการทดลอง แสดงเป็นเปอร์เซ็นต์ของคำที่เกิดขึ้นในแต่ละกรณี | 42 |
| ตารางที่ 4.4 ผลที่ได้จากการทดลอง แสดงเป็นเปอร์เซ็นต์ โดยให้การตั้งคำทุกคำมีค่าเท่ากับ 100 เปอร์เซ็นต์..... | 44 |

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญภาพ

| | หน้า |
|---|------|
| รูปที่ 2.1 ลักษณะของกรณีที่ 1 ประโยคที่เกิดจากคำที่มีในพจนานุกรมทั้งหมด | 7 |
| รูปที่ 2.2 การนำคำที่ดึงได้ มาจัดเรียงให้ได้ประโยคเดิม | 8 |
| รูปที่ 2.3 ลักษณะของกรณีที่ 2 แบบที่ 1 | 10 |
| รูปที่ 2.4 ลักษณะของกรณีที่ 2 แบบที่ 2 | 11 |
| รูปที่ 2.5 ลักษณะของกรณีที่ 2 แบบที่ 3 | 13 |
| รูปที่ 2.6 ลักษณะของกรณีที่ 2 แบบที่ 4 | 14 |
| รูปที่ 2.7 ลักษณะของกรณีที่ 2 แบบที่ 5 | 16 |
| รูปที่ 3.1 แนวความคิดที่ 1 | 18 |
| รูปที่ 3.2 คำที่สามารถดึงออกได้จากข้อความ "ทดสอบภาษาไทย" | 19 |
| รูปที่ 3.3 แนวความคิดที่ 2 | 20 |
| รูปที่ 3.4 คำที่ได้จากการข้อความ "ทดสอบภาษาไทย" | 21 |
| รูปที่ 3.5 แนวความคิดที่ 3 | 22 |
| รูปที่ 3.6 คำที่ได้จากข้อความ "นายเจมส์กระโดด" | 23 |
| รูปที่ 3.7 กราฟการต่อและทับกันของคำจากข้อความ "นายเจมส์มาริตินต้องการผลิตรายการโทรทัศน์" | 30 |
| รูปที่ 3.8 แสดงการนำเอาอัลกอริทึมที่ได้ไปประยุกต์ใช้งาน | 34 |
| รูปที่ 4.1 กราฟแสดงจำนวนเปอร์เซ็นต์ของคำหลัก ที่ได้จากแต่ละขั้นตอนของอัลกอริทึมการจัดทำดัชนี | 41 |
| รูปที่ 4.2 กราฟแสดงจำนวนค่าของน้ำหนักที่ได้จากขั้นตอนที่ 2 ของอัลกอริทึมการจัดทำดัชนี | 43 |
| รูปที่ 4.3 กราฟแสดงจำนวนคำหลักที่ได้จากอัลกอริทึมต่างๆ | 45 |
| รูปที่ ค.1 แสดงตัวอย่าง sistrings | 66 |
| รูปที่ ค.2 ลักษณะโครงสร้างข้อมูลแบบทรี | 67 |