

ระบบการค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพต



นายเปรมิน จินดาวิมลเลิศ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

ภาควิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

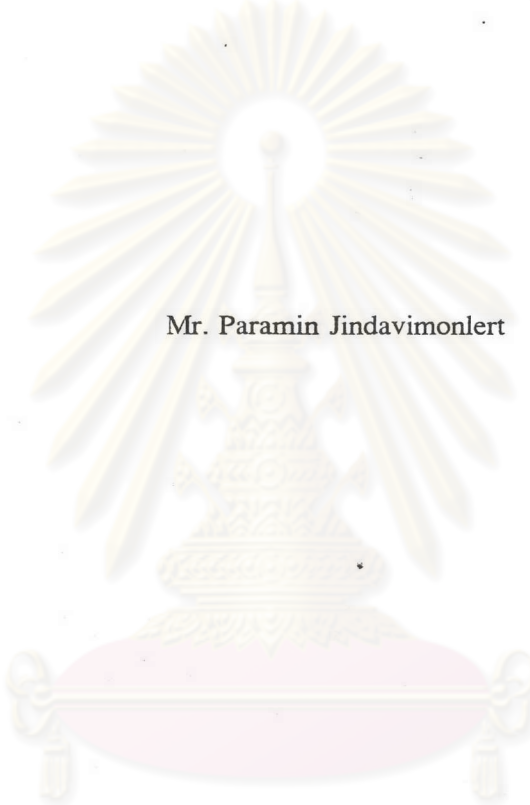
ปีการศึกษา 2539

ISBN 974-636-136-8

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

I17280953

A THAI TEXT RETRIEVAL SYSTEM USING THE PAT TREE



Mr. Paramin Jindavimonlert

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science

Department of Computer Engineering

Graduate School

Chulalongkorn University

Academic Year 1996

ISBN 974-636-136-8

หัวข้อวิทยานิพนธ์

โดย

ภาควิชา

อาจารย์ที่ปรึกษา

ระบบการค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพต

นายเปรมิน จินดาวิมลเลิศ

วิศวกรรมคอมพิวเตอร์

ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จตุระกุล

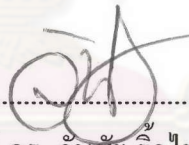
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้หัวข้อวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่ง
ของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต



คณบดีบัณฑิตวิทยาลัย

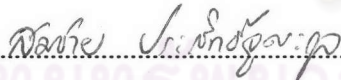
(ศาสตราจารย์ นายแพทย์ สุภวัฒน์ ชุตินวงศ์)

คณะกรรมการสอบวิทยานิพนธ์



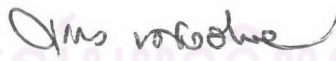
ประธานกรรมการ

(รองศาสตราจารย์ ดร. วันชัย ธีรไพบูลย์)



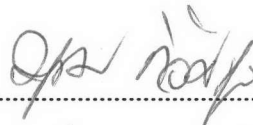
อาจารย์ที่ปรึกษา

(ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จตุระกุล)



กรรมการ

(อาจารย์ ดร. ยรรยง เต็งอำนวย)



กรรมการ

(อาจารย์ ดร. บุญเสริม กิจศิริกุล)



พิมพ์ต้นฉบับบทความวิทยานิพนธ์ภายในกรอบสี่เหลี่ยมนี้เพียงแผ่นเดียว

เปรมิน จินดาวิมลเลิศ : ระบบการค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพ็ค (A THAI TEXT RETRIEVAL SYSTEM USING THE PAT TREE) อ.ที่ปรึกษา : ผศ. ดร. สมชาย ประสิทธิ์จตุระกุล, 78 หน้า. ISBN 974-636-136-8.

วิทยานิพนธ์นี้นำเสนอการพัฒนากระบวนวิธีค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพ็ค เนื่องจากต้นไม้แพ็คจัดเก็บโครงสร้างของสายอักขระแบบกึ่งอนันต์ ที่เรียกว่าซิสตริง ซึ่งคือลำดับย่อยของตัวอักษรต่อเนื่องกันในข้อความ จึงขจัดปัญหาการแบ่งคำในข้อความภาษาไทยที่มักกระทำไม่ได้ไม่ต้องสมบูรณ์ ขั้นตอนการสร้างเริ่มแบ่งเอกสารฉบับใหม่ออกเป็นเอกสารฉบับย่อยๆ ที่มีขนาดแปรตามขนาดของหน่วยความจำหลักที่มีเหลืออยู่ในระบบ จากนั้นตัดซิสตริงที่มีจุดเริ่มต้นที่ไม่ถูกต้องกับหลักภาษาไทยเบื้องต้นออกจากขั้นตอนการจัดเก็บแล้วสร้างแพ็คอะเรย์ที่เก็บซิสตริงที่เหลือของแต่ละเอกสารย่อย โดยใช้ต้นไม้แพ็คเป็นโครงสร้างข้อมูลภายในเมื่อได้แพ็คอะเรย์ของทุกๆ เอกสารย่อยแล้ว จึงนำแพ็คอะเรย์เหล่านี้มาผสานเข้าด้วยกันกับแพ็คอะเรย์เดิมของเอกสารก่อนๆ เป็นแพ็คอะเรย์ใหม่ที่เก็บโครงสร้างของเอกสารฉบับใหม่นั้นด้วย จากนั้นสร้างโครงสร้างระดับที่สอง (ที่มีขนาดเพียงพอต่อการจัดเก็บในหน่วยความจำหลัก) สำหรับอ้างอิงข้อมูลในแพ็คอะเรย์ เพื่อเป็นการเพิ่มประสิทธิภาพการเข้าถึงข้อมูล

ขั้นตอนการสร้างที่ได้กล่าวถึงนี้ใช้เวลาการทำงานเป็น $O(k(N+n))$ โดยที่ k คือจำนวนเอกสารย่อย n คือขนาดของเอกสารฉบับใหม่ที่จะถูกเพิ่ม และ N คือขนาดของแพ็คอะเรย์ก่อนการเพิ่มเอกสารฉบับใหม่ นอกจากนี้ยังได้แสดงให้เห็นว่าแพ็คอะเรย์จะมีขนาดที่แปรผันตามขนาดของซิสตริง จากการทดลองกับข้อความภาษาไทยพบว่า อัตราการเพิ่มขนาดของแพ็คอะเรย์น้อยกว่า 1% เมื่อซิสตริงมีความยาวตั้งแต่ 15 ตัวอักษรเป็นต้นไป

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา วิศวกรรม คอมพิวเตอร์
สาขาวิชา วิทยาการคอมพิวเตอร์
ปีการศึกษา 2539

ลายมือชื่อนิติ เปรมิน จินดาวิมลเลิศ
ลายมือชื่ออาจารย์ที่ปรึกษา ประสิทธิ์จตุระกุล
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสี่เหลี่ยมนี้เพียงแผ่นเดียว

C718342 : MAJOR COMPUTER SCIENCE

KEY WORD: PAT TREE / PAT ARRAY / THAI TEXT RETRIEVAL / TEXT RETRIEVAL

PARAMIN JINDAVIMONLERT : A THAI TEXT RETRIEVAL SYSTEM USING THE
PAT TREE. THESIS ADVISOR : ASSIST. PROF. SOMCHAI PRASITJUTRAKUL, Ph.D.
78 pp. ISBN 974-636-136-8.

This thesis presents a development of Thai text retrieval system using PAT trees. By organizing semi-infinite strings (sistring) as indices in PAT trees which are subsequences of characters, it eliminates the need to perform Thai text segmentation which is usually not 100% correct. The building process begins by first dividing a new document into a set of small subdocuments whose sizes depend on available memory in the system. Next any sistring with ineligible Thai starting characters are eliminated for further consideration. Then a PAT array is built from the set of eligible sistrings by using a PAT trees as one of internal data structures. After obtaining PAT arrays for all the subdocuments, these PAT arrays are then merged with the original PAT array to form a new PAT array having indices for the new document. Finally, a second level index (whose size is sufficiently small for keeping in the available memory) is built for the entire PAT array in order to improve access time.

The whole building process takes time in $O(k(N+n))$ where k is the number of subdocuments, n is the size of the new document to be added, and N is the size of the PAT array before adding the new document. It is also shown that the size of the PAT array increases as sistrings get longer. Experimental results showed that the growth rate of PAT array's size is less than 1% when sistring is of length starting from 15 characters.

ภาควิชา..... วิศวกรรม คอมพิวเตอร์
สาขาวิชา..... วิทยาการคอมพิวเตอร์
ปีการศึกษา..... 2539

ลายมือชื่อนิสิต..... ประมัญ ดินตาอิมล เจริญ
ลายมือชื่ออาจารย์ที่ปรึกษา..... นิตยา ประสงค์สุธา : ศอ
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของ ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จตุระกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่าน ได้สละเวลาให้คำแนะนำ และข้อคิดเห็นต่างๆ ของการวิจัยมาด้วยดีตลอด ขอขอบพระคุณ เจ้าหน้าที่ประจำห้องสมุด และห้องคอมพิวเตอร์ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย ที่ให้บริการศึกษาค้นคว้า เอื้อเพื่ออุปกรณ์บริภัณฑ์ นอกจากนี้ขอ ขอบพระคุณเพื่อนๆ พี่ๆ น้องๆ ชวนนิสิตปริญญาโททุกท่านที่คอยเป็นกำลังใจตลอดมา

ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา-มารดา ซึ่งสนับสนุนในด้านการเงิน รวมทั้งพี่และน้องที่คอยให้กำลังใจแก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา

ประจักษ์ อินทวัฒน์กุล

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญ

	หน้า
บทคัดย่อภาษาไทย.....	ง
บทคัดย่อภาษาอังกฤษ.....	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญตาราง.....	ฅ
สารบัญภาพ.....	ญ

บทที่

1. บทนำ.....	1
1.1 ความเป็นมาและความสำคัญของปัญหา.....	1
1.2 วัตถุประสงค์.....	2
1.3 ขอบเขตของการวิจัย.....	3
1.4 ขั้นตอนและวิธีการดำเนินการวิจัย.....	3
1.5 ประโยชน์ที่คาดว่าจะได้รับ.....	3
1.6 โครงสร้างของวิทยานิพนธ์.....	4
2. แนวคิดและทฤษฎี.....	5
2.1 แนวคิดและทฤษฎี.....	5
2.2 โครงสร้างข้อมูลของระบบการค้นคืน.....	5
2.3 โครงสร้างข้อมูลแบบเพิ่มผกผัน.....	6
2.4 โครงสร้างข้อมูลแบบเพิ่มซิกเนเจอร์.....	7
2.5 โครงสร้างข้อมูลแบบทรี.....	8
2.6 โครงสร้างข้อมูลแบบดับเบิลอะเรย์.....	8
2.7 โครงสร้างข้อมูลแบบต้นไม้แผ่.....	9
3. การวิเคราะห์และการออกแบบ.....	15
3.1 การวิเคราะห์และการออกแบบ.....	15

3.2 ลักษณะเอกสาร.....	15
3.3 การค้นหาข้อความในเอกสาร.....	16
3.4 ดัชนีไม้ค้นหาแบบดิจิทัล.....	17
3.5 โครงสร้างข้อมูลดัชนีไม้แพ็ค.....	19
3.6 การแทนดัชนีไม้แพ็คด้วยอะเรย์.....	23
3.7 การค้นหาในแพ็คอะเรย์.....	24
4. การพัฒนาระบบการค้นคืนข้อความภาษาไทยโดยใช้ดัชนีไม้แพ็ค.....	25
4.1 การพัฒนาระบบการค้นคืนข้อความภาษาไทยโดยใช้ดัชนีไม้แพ็ค.....	25
4.2 ระบบการค้นคืนข้อความภาษาไทย.....	25
4.3 การสร้างแพ็คอะเรย์.....	25
4.3.1 การเตรียมเอกสาร.....	27
4.3.2 การสร้างแพ็คอะเรย์.....	29
4.3.3 การผสานแพ็คอะเรย์.....	34
4.3.4 ตัวอย่างการสร้างแพ็คอะเรย์.....	38
4.4 การค้นคืนข้อความ.....	43
4.4.1 การสร้างแพ็คอะเรย์สั้น.....	45
4.4.2 การค้นคืนแบบเต็มหน้า.....	49
4.4.3 การค้นคืนแบบบูล.....	52
5. การทดสอบความสัมพันธ์ของขนาดของคีย์กับขนาดแฟ้มดรรชนี.....	56
5.1 ตารางแสดงข้อมูลที่ใช้ในการทดสอบ.....	56
5.2 กราฟแสดงความสัมพันธ์.....	57
5.3 แหล่งข้อมูลภาษาไทย.....	58
5.4 การคำนวณหาขนาดแฟ้มดรรชนี.....	58
6. ผลการทดสอบโปรแกรม.....	59
7. บทสรุปและข้อเสนอแนะ.....	65
รายการอ้างอิง.....	66
ภาคผนวก ก.....	68
ประวัติผู้เขียน.....	78

สารบัญตาราง

หน้า

ตารางที่ 4.1 ส่วนหัวเพิ่มเติมตำแหน่งซิสตริง.....	29
ตารางที่ 4.2 ส่วนหัวเพิ่มเติมเพื่อตะเอย์.....	32
ตารางที่ 4.3 ส่วนหัวเพิ่มเติมเพื่อตะเอย์ฉบับใหม่.....	37
ตารางที่ 4.4 โครงสร้างเพิ่มเติมตารางการผสานเพื่อตะเอย์.....	37
ตารางที่ 4.5 ส่วนหัวเพิ่มเติมเพื่อตะเอย์สั้น.....	47
ตารางที่ 5.1 แสดงข้อมูลที่ใช้ในการทดสอบ.....	56
ตารางที่ 6.1 แสดงข้อมูลที่ใช้ในการทดสอบ.....	59
ตารางที่ 6.2 แสดงข้อมูลที่ใช้ในการทดสอบ.....	61
ตารางที่ 6.3 แสดงข้อมูลที่ใช้ในการทดสอบ.....	63
ตารางที่ 6.4 แสดงข้อมูลที่ใช้ในการทดสอบ.....	64

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญภาพ

หน้า

รูปที่ 2.1 ความสัมพันธ์ระหว่างเพิ่มผกผันกับเพิ่มข้อมูล.....	6
รูปที่ 2.2 การคำนวณหาเลขที่อยู่ของข้อมูล.....	7
รูปที่ 2.3 ตัวอย่างโครงสร้างข้อมูลทรี.....	8
รูปที่ 2.4 ลักษณะโครงสร้างข้อมูลแบบดับเบิลอะเรย์.....	9
รูปที่ 2.5 องค์ประกอบของต้นไม้แฝด.....	10
รูปที่ 2.6 ต้นไม้แฝดเมื่อเพิ่มชีสตรงไปแล้ว 8 ชีสตรง.....	12
รูปที่ 2.7 การค้นหาแบบเต็มหน้า.....	13
รูปที่ 3.1 ตัวอย่างต้นไม้ค้นหาแบบดิจิทัล.....	17
รูปที่ 3.2 ตัวอย่างต้นไม้ค้นหาแบบดิจิทัลลดการเปรียบเทียบกรณี.....	18
รูปที่ 3.3 ตัวอย่างต้นไม้แฝด.....	19
รูปที่ 3.4 ฟังก์ชันการค้นหาข้อมูลในต้นไม้แฝด.....	21
รูปที่ 3.5 การเพิ่มข้อมูลที่โหนดภายนอกของต้นไม้แฝด.....	21
รูปที่ 3.6 การเพิ่มข้อมูลที่โหนดภายในของต้นไม้แฝด.....	22
รูปที่ 3.7 ฟังก์ชันการเพิ่มข้อมูลในต้นไม้แฝด.....	23
รูปที่ 3.8 การแทนต้นไม้แฝดด้วยอะเรย์.....	24
รูปที่ 4.1 การสร้างแฟ็ตอะเรย์.....	26
รูปที่ 4.2 ต้นไม้แฝดที่มีชีสตรงทั้งหมด 9 ชีสตรงและชีสตรงซ้ำๆ กัน 3 ชีสตรง.....	30
รูปที่ 4.3 โครงสร้างโหนดต้นไม้แฝด.....	31
รูปที่ 4.4 โครงสร้างแฟ็ตอะเรย์.....	33
รูปที่ 4.5 โครงสร้างตัวนับค่าของแฟ็ตอะเรย์ย่อย.....	35
รูปที่ 4.6 โครงสร้างตารางการผสมแฟ็ตอะเรย์.....	36
รูปที่ 4.7 ตัวอย่างการเตรียมเอกสาร.....	39
รูปที่ 4.8 ตัวอย่างการสร้างแฟ็ตอะเรย์.....	40
รูปที่ 4.9 ตัวอย่างการผสมแฟ็ตอะเรย์.....	42

รูปที่ 4.11 ตัวอย่างการสร้างแพ็คเกจเรย์สั้น.....	47
รูปที่ 4.12 ตัวอย่างการคั่นคืนแบบเต็มหน้า.....	51
รูปที่ 4.13 ตัวอย่างการคั่นคืนแบบบุล.....	54
รูปที่ 5.1 แสดงความสัมพันธ์ระหว่างขนาดเพิ่มครรชนีกับขนาดของคีย์.....	57
รูปที่ 6.1 แสดงความสัมพันธ์ระหว่าง เวลาในการสร้างครรชนีกับขนาดเอกสาร.....	60
รูปที่ 6.2 แสดงความสัมพันธ์ระหว่าง เวลาในการสร้างครรชนีกับความยาวซิสตริง.....	62
รูปที่ 6.3 แสดงความสัมพันธ์ระหว่าง เวลาในการสร้างครรชนีกับขนาดเอกสารย่อย.....	63
รูปที่ 6.4 แสดงความสัมพันธ์ระหว่าง เวลาในการสร้างครรชนีกับขนาดเอกสารและขนาดเอกสารย่อย.....	64

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย