ระบบการค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพ็ต

นายเปรมิน จินดาวิมลเลิศ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
ภาควิชาวิศวกรรมคอมพิวเตอร์
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2539

ISBN 974-636-136-8

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	ระบบการค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพ็ต
โดย	นายเปรมิน จินดาวิมลเลิศ
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. ยรรยง เต็งอำนวย)

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

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสีเขียวนี้เพียงแผ่นเดียว

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

วิทยานิพนธ์นี้นำเสนอการพัฒนาระบบค้นคืนข้อความภาษาไทยโดยใช้ต้นไม้แพ็ต เนื่องจากต้นไม้แพ็ตจัดเก็บดรรชนีของสายอักขระแบบกึ่งอนันต์ ที่เรียกว่าซิสตริง ซึ่งคือลำดับย่อยของตัวอักษรต่อเนื่องกันในข้อความ จึงขจัดปัญหาการแบ่งคำในข้อความภาษาไทยที่มักกระทำได้ไม่ถูกต้องสมบูรณ์ ขั้นตอนการสร้างเริ่มแบ่งเอกสารฉบับใหม่ออกเป็นเอกสารฉบับย่อยๆ ที่มีขนาดแปรตามขนาดของหน่วยความจำหลักที่มีเหลืออยู่ในระบบ จากนั้นตัดซิสตริงที่มีจุดเริ่มต้นที่ไม่ถูกต้องกับหลักภาษาไทยเบื้องต้นออกจากขั้นตอนการจัดเก็บแล้วสร้างแพ็ตอะเรย์ที่เก็บซิสตริงที่เหลือของแต่ละเอกสารย่อย โดยใช้ต้นไม้แพ็ตเป็นโครงสร้างข้อมูลภายใน เมื่อได้แพ็ตอะเรย์ของทุกๆ เอกสารย่อยแล้ว จึงนำแพ็ตอะเรย์เหล่านี้มาผสานเข้าด้วยกันกับแพ็ตอะเรย์เดิมของเอกสารก่อนๆ เป็นแพ็ตอะเรย์ใหม่ที่เก็บดรรชนีของเอกสารฉบับใหม่นั้นด้วย จากนั้นสร้างดรรชนีระดับที่สอง (ที่มีขนาดเพียงพอต่อการจัดเก็บในหน่วยความจำหลัก) สำหรับอ้างอิงข้อมูลในแพ็ตอะเรย์ เพื่อเป็นการเพิ่มประสิทธิภาพการเข้าถึงข้อมูล

ขั้นตอนการสร้างที่ได้กล่าวถึงนี้ใช้เวลาการทำงานเป็น $O(k(N+n))$ โดยที่ k คือจำนวนเอกสารย่อย n คือขนาดของเอกสารฉบับใหม่ที่จะถูกเพิ่ม และ N คือขนาดของแพ็ตอะเรย์ก่อนการเพิ่มเอกสารฉบับใหม่ นอกจากนี้ยังได้แสดงให้เห็นว่าแพ็ตอะเรย์จะมีขนาดที่แปรผันตามขนาดของซิสตริง จากการทดลองกับข้อความภาษาไทยพบว่า อัตราการเพิ่มขนาดของแพ็ตอะเรย์น้อยกว่า 1% เมื่อซิสตริงมีความยาวตั้งแต่ 15 ตัวอักษรเป็นต้นไป

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต เปรมิน จินดาวิมลเลิศ

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา สมชาย ประสิทธิ์จูตระกูล

ปีการศึกษา 2539 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม .....

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสีเขียวนี้เพียงแผ่นเดียว

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 futher 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 วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต ปรมิน จินดาวิมลเลิศ

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา สมชาย ประสิทธิ์จุตระกูล

ปีการศึกษา 2539 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของ ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จูตระกูล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้สละเวลาให้คำแนะนำ และข้อคิดเห็นต่างๆ ของการวิจัยมาด้วยดีตลอด ขอขอบพระคุณเจ้าหน้าที่ประจำห้องสมุด และห้องคอมพิวเตอร์ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย ที่ให้บริการศึกษาค้นคว้า เอื้อเฟื้ออุปกรณ์บริภัณฑ์ นอกจากนี้ขอขอบพระคุณเพื่อนๆ พี่ๆ น้องๆ ชาวนิสิตปริญญาโททุกท่านที่คอยเป็นกำลังใจตลอดมา

ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา-มารดา ซึ่งสนับสนุนในด้านการเงิน รวมทั้งพี่และน้องที่คอยให้กำลังใจแก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา

เปรมิน ลิ้มตระมูลชัย

สารบัญ

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า