การพัฒนาโครงสร้างข้อมูลพจนานุกรมไทยด้วยทรัยแถวคู่

นาย อภิชิต พิทยรัตน์โสภณ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
ภาควิชาวิศวกรรมคอมพิวเตอร์
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
พ.ศ. 2537

ISBN 974-584-532-9

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

I16923911

THAI DICTIONARY DATA STRUCTURE DEVELOPMENT USING DOUBLE ARRAY TRIE

Mr. Apichit Pittayaratsopon

A Thesis Submitted in Partial Fulfillment of the Requirement

for the Degree of Master of Science

Department of Computer Engineering

Graduate School

Chulalongkorn University

1994

ISBN 974-584-532-9

หอสมุดกลาง สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

หัวข้อวิทยานิพนธ์	การพัฒนาโครงสร้างข้อมูลพจนานุกรมไทยด้วยทรัยแถวคู่
โดย	นาย อภิชิต พิทยรัตน์โสภณ
ภาควิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	อาจารย์ ดร. ยรรยง เต็งอำนวย
อาจารย์ที่ปรึกษาร่วม	อาจารย์ วิรัช ศรเลิศล้ำวาณิช

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีบัณฑิตวิทยาลัย

(ศาสตราจารย์ ดร. ถาวร วัชราภัย)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(รองศาสตราจารย์ สมชาย ทยานยง)

.....กรรมการ

(อาจารย์ ดร. ยรรยง เต็งอำนวย)

.....กรรมการ

(อาจารย์ วิรัช ศรเลิศล้ำวาณิช)

.....กรรมการ

(อาจารย์ จารุมาตร ปิ่นทอง)

.....กรรมการ

(อาจารย์ ดร. สมชาย ประสิทธิ์จูตระกูล)

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสีเขียวนี้เพียงแผ่นเดียว

อภิชิต พิทยรัตน์โสภณ : การพัฒนาโครงสร้างข้อมูลพจนานุกรมไทยด้วยทรัยแถวคู่ (THAI DICTIONARY DATA STRUCTURE DEVELOPMENT USING DOUBLE ARRAY TRIE) อ.ที่ปรึกษา : อ.ดร. ยรรยง เด็งอำนวย, อ.ที่ปรึกษาร่วม : อ. วิรัช ศรเลิศล้ำวาณิช, 53 หน้า, ISBN 974-584-532-9

การวิจัยครั้งนี้มีจุดมุ่งหมายเพื่อสร้างและพัฒนาโครงสร้างข้อมูลของระบบจัดการฐานข้อมูลพจนานุกรมภาษาไทยที่มีลักษณะจลน์ สามารถเพิ่มลบคำได้ตลอดโดยไม่ต้องสร้างพจนานุกรมทั้งหมดใหม่ เพื่อใช้ในโครงการแปลภาษาด้วยคอมพิวเตอร์ โดยให้มีประสิทธิภาพในการสืบค้นเพิ่มลบคำศัพท์ ทั้งในด้านความเร็ว ความยืดหยุ่น รวมถึงการประหยัดหน่วยความจำ ด้วยทรัยแถวคู่

มีการใช้โครงสร้างข้อมูลต่างๆเพื่อเก็บพจนานุกรม รวมทั้งโครงสร้างข้อมูลทรัยที่เข้ากันได้ดีกับลักษณะของคำไทย แต่โครงสร้างทั้งหมดล้วนมีลักษณะสถิตย์ คือไม่สามารถเพิ่มลบคำได้ในทันที ต้องแก้ไขเพิ่มลบคำนั้นในชุดของคำศัพท์ก่อน แล้วจึงสร้างพจนานุกรมทั้งหมดขึ้นมาใหม่ แต่เมื่อใช้โครงสร้างข้อมูลทรัยแถวคู่ซึ่งมีลักษณะจลน์แล้วสามารถลบจุดบกพร่องนี้ได้

จากผลการทดสอบพบว่าทรัยแถวคู่เหมาะที่จะใช้เป็นโครงสร้างข้อมูลสำหรับฐานข้อมูลพจนานุกรมภาษาไทยที่จำเป็นต้องมีการแก้ไขเพิ่มลบคำอยู่ตลอด เพราะว่ามีความสามารถในการเพิ่มลบคำได้ มีประสิทธิภาพการใช้เนื้อที่หน่วยความจำที่ดี และมีเวลาที่ใช้ในการเพิ่มลบคำที่ดี

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา .....

ปีการศึกษา 2536 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม .....

C216583 : MAJOR COMPUTER SCIENCE
KEY WORD: DATA STRUCTURE / DICTIONARY / THAI LANGUAGE / DOUBLE ARRAY TRIE
APICHIT PITTAYARATSOPON : THAI DICTIONARY DATA STRUCTURE USING DOUBLE ARRAY TRIE. THESIS ADVISOR : YANYONG TENGAMNUAY, Ph.D., THESIS CO-ADVISOR : VIRACH SORNLERTLAMVANICH. 53 pp. ISBN 974-584-532-9

This research has objective to create and develop data structure of Thai dictionary database management system which has dynamic characteristic and will be used in machine translation project. Insertion and deletion of words can be done as often as needed without creating new whole dictionary. The development uses Double Array Trie for efficiencies of insert and delete time, flexibility, and memory use.

There are various data structures used to develop dictionary, including trie data structure which matches Thai word characteristic. But all are static - words cannot be inserted or deleted immediately, but words are initially inserted or deleted in set of words and then new dictionary is recreated. When using dynamic Double Array Trie data structure, above problem can be solved.

From result of testing, Double Array Trie is suitable to be used as data structure of Thai dictionary database which is necessary to have words edited as often as needed. Because Double Array Trie has ability of insertion and deletion, efficiency in memory using, and fast insert and delete time.

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา .....

ปีการศึกษา 2536 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม .....

หอสมุดกลาง สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลือจากบุคคลหลายท่าน ผู้เขียนขอขอบพระคุณอาจารย์ ดร. ยรรยง เต็งอำนวยสำหรับคำแนะนำปรึกษา ขอขอบพระคุณอาจารย์ วิรัช ศรเลิศ-ล้ำวาณิช อาจารย์ ดร. สุรพันธ์ เมฆนาวิน และบุคลากรห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้ ศูนย์เทคโนโลยีอิเลกทรอนิกส์และคอมพิวเตอร์แห่งชาติสำหรับข้อมูล เครื่องมือ และสถานที่ และขอขอบคุณเพื่อนพี่น้องสำหรับคำแนะนำต่างๆ

ท้ายนี้ ผู้เขียนขอขอบพระคุณบิดามารดาที่ให้กำลังใจเสมอมาจนสำเร็จการศึกษา

สารบัญ

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญตาราง ..... ฌ

สารบัญภาพ ..... ญ

บทนำ 1

- ความสำคัญ และความเป็นมาของปัญหา 1
  - โครงสร้างข้อมูล 2
  - โครงสร้างทรัย 3
- วัตถุประสงค์ของการวิจัย 7
- ขอบเขตของการวิจัย 7
- วิธีดำเนินการวิจัย 8
- ประโยชน์ที่คาดว่าจะได้รับ 8

โครงสร้างข้อมูลทรัยแถวคู่ 10

- การเปรียบแบบ (Pattern Matching) ของทรัยแถวคู่ 10
- การปรับปรุงทรัยแถวคู่ 13
  - เงื่อนไขเพิ่มเติมของทรัยแถวคู่ 13
  - การสืบค้นคีย์ 15
  - การเพิ่มคีย์ 17
  - การลบคีย์ 23

การออกแบบโครงสร้างข้อมูลทรัยแถวคู่ 25

- โครงสร้างข้อมูลของอาร์เรย์คู่ (BASE และ CHECK) 25
- โครงสร้างแฟ้มข้อมูลของอาร์เรย์คู่ (BASE และ CHECK) 26

- โครงสร้างแฟ้มข้อมูลของ TAIL 27
- ค่าคงที่ที่ใช้ในโปรแกรม 27
- โครงสร้างข้อมูลที่ใช้ในโปรแกรม 29
- ฟังก์ชั่นการจัดการโครงสร้างข้อมูลทรัยแถวคู่ 31
- โปรแกรมอรรถประโยชน์ 33

การทดสอบ 35
- โครงสร้างข้อมูลทรัยอัดแน่นแบบจลน์ 35
- ผลการทดสอบ 36
- สรุปผล 37

บทสรุปและข้อเสนอแนะ 40
- บทสรุป 40
- ข้อเสนอแนะ 41

เอกสารอ้างอิง 42

ประวัติผู้เขียน 43

สารบัญตาราง

ตารางที่ 4.1 ผลการทดสอบกับคำประเภทต่างๆ 39

สารบัญรูป

รูปที่ 1.1 คำภาษาอังกฤษที่ใช้บ่อย 31 คำ 3
รูปที่ 1.2 ทรัยโยงของคำอังกฤษที่ใช้บ่อย 4
รูปที่ 1.3 ทรัยดรรชนีของคำอังกฤษที่ใช้บ่อย 4
รูปที่ 1.4 ทรัยอัดแน่นของคำอังกฤษที่ใช้บ่อย 5
รูปที่ 2.1 เครื่องเปรียบแบบ (Pattern Matching Machine) ของ K1 11
รูปที่ 2.2 โครงสร้างข้อมูลทรัยของ K1 12
รูปที่ 2.3 โครงสร้างข้อมูลทรัยแถวคู่ 12
รูปที่ 2.4 ฟังก์ชั่น goto ของ K' 13
รูปที่ 2.5 ทรัยแถวคู่ของ K' 16
รูปที่ 2.6 ขั้นตอนการเพิ่มคีย์ 21
รูปที่ 3.1 แสดง BASE และ TAIL ที่ปรับปรุงแล้ว 26
รูปที่ 3.2 แสดง BASE และ CHECK ที่เก็บเป็นแฟ้มข้อมูล 26
รูปที่ 3.3 แสดง TAIL ที่เก็บเป็นแฟ้มข้อมูล 27