

โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย

นางสาวสมปราวณา รัชยานนท์



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

ภาควิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

พ.ศ. 2535

ISBN 974-581-176-9

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

018719

117146343

DATA STRUCTURE FOR THAI ELECTRONIC DICTIONARY

Miss Somprathana Ratthayanond

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science
Department of Computer Engineering
Graduate School

1992

ISBN 974-581-176-9

หัวข้อวิทยานิพนธ์
โดย
ภาควิชา
อาจารย์ที่ปรึกษา

โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย
นางสาวสมปราวรณา รัชยานนท์
วิศวกรรมคอมพิวเตอร์
รองศาสตราจารย์ ดร. วิลาศ วุวงศ์
ผู้ช่วยศาสตราจารย์ สุธุชน์ สัตยประกอบ

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้วิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่ง
ของการศึกษาตามหลักสูตรปริญญาโทบัณฑิต

..... คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ ดร. ดาวร วัชรไวย)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รศ. เดือน สิ้นสุนันต์ประทุม)

..... อาจารย์ที่ปรึกษา
(รศ.ดร. วิลาศ วุวงศ์)

..... อาจารย์ที่ปรึกษาร่วม
(ผศ. สุธุชน์ สัตยประกอบ)

..... กรรมการ
(อ.ดร. ยรรยง เต็งอำนวย)

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสี่เหลี่ยมนี้เพียงแผ่นเดียว



เล่มปราศรัย วิทยานิพนธ์ : โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย
(DATA STRUCTURE FOR THAI ELECTRONIC DICTIONARY) อ.ที่ปรึกษา : รศ.ดร.วิลาศ
ววงค์, ผศ.สุยุษณ์ สัตยประกอบ, 90 หน้า. ISBN 974-581-176-9.

พจนานุกรมอิเล็กทรอนิกส์เป็นแหล่งเก็บข้อมูลสำหรับงานประมวลผลด้านภาษาคำศัพท์ เช่น การตัดคำและการตรวจสอบตัวสะกดในโปรแกรมประมวลผลคำ การวิเคราะห์ไวยากรณ์ในงานประมวลผลภาษาธรรมชาติ จากความก้าวหน้าของงานการประมวลผล ภาษาไทยด้วยคอมพิวเตอร์ที่มีมากขึ้น ทำให้มีความสนใจในการพัฒนาพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยมากขึ้นตามไปด้วย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยโดยใช้โครงสร้างข้อมูลแบบดับเบิลเอเรย์ที่มีการสืบค้นแบบดิคชันนารี 2 แบบ ดังนี้คือ แบบแรก เป็นพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่จัดเก็บคำศัพท์ต่าง ๆ โดยตรง ส่วนแบบที่ 2 เป็นพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่จัดเก็บคำโดดที่ได้มาจากการแยกคำศัพท์ ซึ่งผลของการจัดเก็บได้ว่าเนื้อที่ที่ใช้ในการเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยแบบที่ 2 น้อยกว่าแบบแรก นอกจากนี้ ยังพัฒนาอัลกอริทึมการตัดคำโดยใช้พจนานุกรมที่ให้ผลลัพธ์เป็นคำศัพท์ทุกคำที่ปรากฏในพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย

สำหรับผลการทดสอบประสิทธิภาพการทำงานพบว่า อัลกอริทึมการสืบค้นคำศัพท์ของพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่พัฒนาขึ้นใช้เวลามากกว่าพจนานุกรมของมหาวิทยาลัยเกษตรศาสตร์ แต่สามารถนำอัลกอริทึมดังกล่าวมาเป็นแนวทางในการพัฒนาอัลกอริทึมการตัดคำโดยใช้พจนานุกรมได้โดยง่าย

ภาควิชา ศึกษาศาสตร์ มหาวิทยาลัยราชภัฏวชิรเวศน์
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์
ปีการศึกษา 2535

ลายมือชื่อนิติบัตร วิชาภาษาไทย วิทยานิพนธ์
ลายมือชื่ออาจารย์ที่ปรึกษา วิลาศ ววงค์
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม สุยุษณ์ สัตยประกอบ

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสี่เหลี่ยมนี้เพียงแผ่นเดียว

##C116994 : MAJOR COMPUTER SCIENCES

KEY WORD : DATA STRUCTURE/Dictionary/ELECTRONIC/THAI

SOMPRATHANA RATTAYANOND : DATA STRUCTURE FOR THAI ELECTRONIC DICTIONARY. THESIS ADVISOR : ASSOCIATE PROF.DR. VILAS WUWONGSE ASSISTANT PROF. SUYUT SATAYAPRAKORB. 90 pp., ISBN 974-581-176-9.

Electronic dictionaries are storages of machine readable lexical items. They are used in advanced word processors for word segmentation and spelling checkers, and natural language processing for syntactic, semantic and discourse analysis. Due to the advancement in computer processing of Thai language, more attention has recently been paid to the development of Thai electronic dictionaries.

This study proposes and develops two frameworks for Thai electronic dictionaries employing double-array digital search tree. The first framework treats a Thai word as a lexical entity and directly applies the double-array digital search tree to store lexical entities. The second one recognizes the fact that many Thai words are composed of few isolated words, and therefore stores a word in terms of a few isolated words, resulting in less storage space. Based on the dictionaries, a Thai word segmentation algorithm has been developed which produces all possible segmentations.

Experiments have been conducted to evaluate performance of the proposed frameworks. It has been found out that it takes more time for the frameworks to retrieve words than the Kasetsart University's approach but they allow the development of simple word segmentation algorithm.

ภาควิชา วิศวกรรมคอมพิวเตอร์

สาขาวิชา วิทยาการคอมพิวเตอร์

ปีการศึกษา 2534

ลายมือชื่อนิติกร สิบประภม รัตนานนท์

ลายมือชื่ออาจารย์ที่ปรึกษา วรวิทย์

ลายมือชื่ออาจารย์ที่ปรึกษาร่วม สุยุต สัตยาพรกอร์บ

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของ
รศ.ดร.วิลาส วรงค์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ และ ผศ.สุยุชน์ สัตยประกอบ
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ซึ่งท่านได้สละเวลาให้คำแนะนำ และข้อคิดเห็นต่างๆ
ของการวิจัยมาด้วยดีตลอด

ขอขอบพระคุณ รศ.ยีน ภู่วรรณ ที่ได้ อนุเคราะห์ข้อมูลสำหรับการทำ
วิทยานิพนธ์ ขอขอบพระคุณ คุณพรพจน์ สิริยวงศ์ ที่เสียสละเวลาให้คำแนะนำ และ
ข้อคิดเห็นต่างๆ ในการทำวิทยานิพนธ์ครั้งนี้ ขอขอบพระคุณเจ้าหน้าที่ประจำห้องปฏิบัติการ
คอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ที่เอื้อเฟื้ออุปกรณ์บริภัณฑ์
และเนื่องจากทุนการวิจัยครั้งนี้บางส่วนได้รับมาจากทุนอุดหนุนการวิจัยของบัณฑิตวิทยาลัย
จึงขอขอบคุณบัณฑิตวิทยาลัยมา ณ ที่นี้ด้วย นอกจากนี้ขอขอบพระคุณเพื่อนๆ พี่ๆ น้องๆ
ชาวนิสิตปริญญาโททุกๆ ท่านที่คอยเป็นกำลังใจตลอดมา

ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา-มารดา ซึ่งสนับสนุนในด้านการเงิน
รวม ทั้ง พี่ และน้องที่คอยให้กำลังใจแก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	จ
กิตติกรรมประกาศ	ฉ
สารบัญตาราง	ญ
สารบัญภาพ	ฎ
บทที่	
1. บทนำ	1
- ความเป็นมาของปัญหา	1
- วัตถุประสงค์ของวิทยานิพนธ์	2
- ขอบเขตและเงื่อนไขของวิทยานิพนธ์	3
- ขั้นตอนการวิจัย	3
- ประโยชน์ที่คาดว่าจะได้รับจากวิทยานิพนธ์	3
- โครงสร้างของวิทยานิพนธ์	4
2. พจนานุกรมอิเล็กทรอนิกส์ภาษาไทย	6
- ประเภทพจนานุกรมอิเล็กทรอนิกส์	6
- รายละเอียดข้อมูลในพจนานุกรมอิเล็กทรอนิกส์	7
- โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์	9
- ผลงานการสร้างพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย ที่สนับสนุนการประมวลผลภาษาไทยด้วย เครื่องคอมพิวเตอร์	17

3. แนวทางการออกแบบพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย	22
- ลักษณะคำไทย	22
- การค้นหาคำศัพท์จากพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย	23
- วิธีปฏิบัติการกับข้อมูล	24
- โครงสร้างข้อมูลแบบดีเอส-ทรี	25
- ลักษณะการเก็บข้อมูลด้วยโครงสร้างข้อมูล แบบดับเบิลโอเชอรี	26
- อัลกอริทึมการสืบค้นข้อมูล	32
- ขั้นตอนวิธีการปรับทันกาลของ โครงสร้างข้อมูลแบบดับเบิลโอเชอรี	36
- สรุปแนวทางการออกแบบโครงสร้างข้อมูล สำหรับเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย	41
4. รายละเอียดการออกแบบและการพัฒนา โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย	42
- ลักษณะโครงสร้างข้อมูลที่ศึกษาจากงานวิจัย	42
- รูปแบบโครงสร้างข้อมูลสำหรับพจนานุกรม อิเล็กทรอนิกส์ภาษาไทย	43
- อัลกอริทึมการสืบค้นคำศัพท์	47
- อัลกอริทึมการเพิ่มคำศัพท์	54
- การพัฒนาโครงสร้างข้อมูลสำหรับจัดเก็บ พจนานุกรมอิเล็กทรอนิกส์ภาษาไทย	61
5. การตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์ภาษาไทย	63
- อัลกอริทึมการตัดคำโดยใช้ พจนานุกรมอิเล็กทรอนิกส์ภาษาไทย	63
- การพัฒนาโปรแกรมการตัดคำ	69

6. การวิเคราะห์ผลการทำงาน	70
- ข้อมูลที่ใช้ในการทดสอบ	70
- ขั้นตอนการทดสอบ	70
7. บทสรุป	75
- ประสิทธิภาพการสืบค้นคำค้นที่ ของพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย	75
- ประสิทธิภาพการตัดคำโดยใช้ พจนานุกรมอิเล็กทรอนิกส์ภาษาไทย	76
- ข้อเสนอแนะ	77
เอกสารอ้างอิง	78
ภาคผนวก ก	81
ภาคผนวก ข	84
ภาคผนวก ค	86
ประวัติผู้เขียน	90

สารบัญตาราง

หน้า

ตารางที่ 2.1	ตารางแสดงการเปรียบเทียบประสิทธิภาพ ของโครงสร้างข้อมูลแบบต่าง ๆ	21
ตารางที่ 3.1	ขั้นตอนการสืบค้นคำศัพท์.....	35
ตารางที่ 4.1	ตัวอย่างการสืบค้นคำศัพท์จากพจนานุกรมอิเล็กทรอนิกส์ ภาษาไทย.....	52
ตารางที่ 4.2	ตัวอย่างการเพิ่มคำศัพท์ในพจนานุกรมอิเล็กทรอนิกส์ ภาษาไทย.....	58
ตารางที่ 5.1	ตัวอย่างการตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์ ภาษาไทย.....	69

สารบัญภาพ

	หน้า
รูปที่ 2.1	โครงสร้างข้อมูลแบบอินเด็กซ์ซีเควนเซียล..... 11
รูปที่ 2.2	โครงสร้างข้อมูลแบบบีทรี..... 13
รูปที่ 2.3	ลักษณะโหนดของโครงสร้างข้อมูลแบบบีทรี..... 13
รูปที่ 2.4	ตัวอย่างการเพิ่มเลข 22 ในบี-ทรี..... 14
รูปที่ 2.5	โครงสร้างข้อมูลแบบทรี..... 16
รูปที่ 2.1	โครงสร้างข้อมูลแบบทีทีเอส..... 18
รูปที่ 2.2	โครงสร้างข้อมูลแบบทีไอเอส..... 19
รูปที่ 2.3	โครงสร้างข้อมูลแบบต้นไม้..... 20
รูปที่ 3.1	โครงสร้างแบบดีเอส-ทรี..... 25
รูปที่ 3.2	ลักษณะโครงสร้างข้อมูลแบบดับเบิลเอเรย์..... 27
รูปที่ 3.3	ความสัมพันธ์ของโครงสร้างแบบดับเบิลเอเรย์..... 30
รูปที่ 3.4	ดีเอส-ทรีของคำศัพท์ตัวอย่าง..... 31
รูปที่ 3.5	ดับเบิลเอเรย์และเทล..... 31
รูปที่ 4.1	แสดงความสัมพันธ์ระหว่างอะเรย์เบส อะเรย์ชี้คและอะเรย์เทล..... 44
รูปที่ 4.2	แสดงรูปแบบการเก็บข้อมูลของโครงสร้างข้อมูล สำหรับเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย..... 45
รูปที่ 4.3	แสดงโหนดต่างๆ ของโครงสร้างข้อมูล สำหรับเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย..... 46
รูปที่ 4.4	แสดงความสัมพันธ์ของอะเรย์เบส อะเรย์ชี้ค และเทล..... 60
รูปที่ 4.5	แสดงลักษณะโครงสร้างพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย..... 60
รูปที่ 5.1	แสดงลักษณะโครงสร้างของประโยค..... 69