

การปรับเปลี่ยนการอัดข้อความภาษาไทย



นางสาวเรวดี ลิ้มปิไชติกุล

ศูนย์วิทยทรัพยากร  
จุฬาลงกรณ์มหาวิทยาลัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

ภาควิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

พ.ศ. 2534


ISBN 974-578-965-8

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

017624

117844712

IMPROVEMENT OF THAI TEXT COMPRESSION



Miss Rawedee Limpichotikul

ศูนย์วิทยทรัพยากร  
จุฬาลงกรณ์มหาวิทยาลัย

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

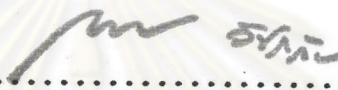
1991

ISBN 974-578-965-8

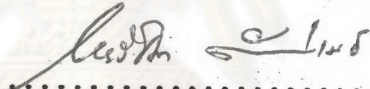


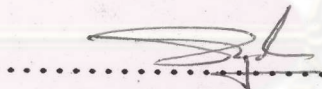
หัวข้อวิทยานิพนธ์ การปรับเปลี่ยนการถอดข้อความภาษาไทย  
โดย นางสาวเรวดี ลิ้มปิไชติกุล  
ภาควิชา วิศวกรรมคอมพิวเตอร์  
อาจารย์ที่ปรึกษา อาจารย์ จารุมาตร ปิ่นทอง  
ผู้ช่วยศาสตราจารย์ ดร. วีระ ธีวพิทักษ์

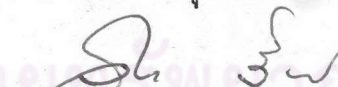
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้บัณฑิตวิทยาลัยนี้ เป็นส่วนหนึ่งของ  
การศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต


  
..... คณบดีบัณฑิตวิทยาลัย  
(ศาสตราจารย์ ดร. ถาวร ธีวราษฎร์)

คณะกรรมการสอบวิทยานิพนธ์

  
..... ประธานกรรมการ  
(รองศาสตราจารย์ ไกรวิจิต ตันติเมธ)

  
..... อาจารย์ที่ปรึกษา  
(อาจารย์ จารุมาตร ปิ่นทอง)

  
..... อาจารย์ที่ปรึกษาร่วม  
(ผู้ช่วยศาสตราจารย์ ดร. วีระ ธีวพิทักษ์)

  
..... กรรมการ  
(รองศาสตราจารย์ เตือน สินธุ์พันธ์ประทุม)

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสี่เหลี่ยมนี้เพียงแผ่นเดียว

เรวดี ลิ้มปิโชติกุล : การปรับเปลี่ยนการอัดข้อความภาษาไทย (IMPROVEMENT OF THAI TEXT COMPRESSION) อ.ที่ปรึกษา : อ.จารุมাত্র ปิ่นทอง อ.ที่ปรึกษาร่วม : ผศ.ดร. วีระ รวีพิทักษ์, 82 หน้า. ISBN 974-578-965-8

การวิจัยครั้งนี้มีจุดมุ่งหมายเพื่อศึกษาหาวิธีการอัดข้อมูลที่มีประสิทธิภาพ และปรับเปลี่ยนวิธีการจากเดิมให้มีประสิทธิภาพเพิ่มขึ้น สำหรับนำไปใช้ในการอัดข้อความภาษาไทย จากการศึกษาการอัดข้อมูลวิธีต่าง ๆ ในปัจจุบัน พบว่าวิธีการที่มีประสิทธิภาพ และเหมาะสมกับการอัดข้อความภาษาไทย ได้แก่ วิธีการของอัลกอริทึมฮัฟแมน (Huffman Algorithm) และวิธีการของอัลกอริทึมแอลแซดดับเบิลว (LZW Algorithm) การปรับเปลี่ยนวิธีการจะนำข้อมูลคำไทยที่พบบ่อยในชีวิตประจำวัน จากงานวิจัยของมหาวิทยาลัยเกษตรศาสตร์ เรื่อง การวิเคราะห์คำไทย (Thai Word Analysis) สร้างเป็นแฟ้มข้อมูลคำไทยเพื่อนำมาใช้ในวิธีการทั้งสอง

ในการทดสอบเพื่อดูประสิทธิภาพการอัดข้อมูลของวิธีการที่ปรับเปลี่ยนกับข้อมูลขนาดต่าง ๆ พบว่าวิธีการที่ปรับเปลี่ยนสามารถให้ประสิทธิภาพการอัดข้อมูลสูงขึ้นกว่าเดิม สามารถลดขนาดข้อมูลลงได้โดยเฉลี่ย 40-55 เปอร์เซ็นต์ ซึ่งขึ้นอยู่กับขนาดของข้อมูลและลักษณะของข้อมูลที่สอดคล้องกับแฟ้มข้อมูลคำไทย เมื่อพิจารณาเปรียบเทียบระหว่างวิธีการที่ปรับเปลี่ยนทั้งสองวิธีพบว่า วิธีการของอัลกอริทึมฮัฟแมนที่ปรับเปลี่ยนเป็นวิธีการที่มีอัลกอริทึมที่ซับซ้อนกว่า และใช้เวลาในการอัดข้อมูลมากกว่าวิธีการของอัลกอริทึมแอลแซดดับเบิลวที่ปรับเปลี่ยน ดังนั้นวิธีการของอัลกอริทึมแอลแซดดับเบิลวที่ปรับเปลี่ยน จึงเป็นวิธีการที่เหมาะสมสำหรับนำไปใช้ในการอัดข้อความภาษาไทย



ภาควิชา วิทยาการคอมพิวเตอร์ ๒๓๐๖  
สาขาวิชา วิทยาการคอมพิวเตอร์ ๒๓๐๖  
ปีการศึกษา ๒๕๓๓

ลายมือชื่อนิสิต เรวดี ลิ้มปิโชติกุล  
ลายมือชื่ออาจารย์ที่ปรึกษา [Signature]  
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม [Signature]

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสี่เหลี่ยมนี้เพียงฉบับเดียว

RAWEDDEE LIMPICHOTIKUL : IMPROVEMENT OF THAI TEXT COMPRESSION.  
THESIS ADVISOR : MR. JARUMATR PINTHONG. ASST.PROF. WEERA  
RIEWPIITUK, Ph.D. 82 pp. ISBN 974-578-965-8

An objective of this research is to search for the efficient methods of data compression and improve the methods to increase the efficiency when compressing Thai text. The two efficient methods, Huffman algorithm and LZW algorithm have been found to gave a high compression efficiency and suitable for use with Thai text. An further improvement, words that are frequently found in Thai text which is obtained from Thai Word Analysis, a research of Kasetsart University, will be used to create a Thai word file for use with these two methods.

From the experiment, examples of various file sizes are presented to demonstrate the increasing performance of improved methods, give a compression efficiency of about 40-55 percents. The increasing performance depending on file sizes and Thai semantics in files. When comparing the improved methods in terms of compression time and algorithm complexity, it show that improved Huffman algorithm take more time than improved LZW algorithm because it is more complex. Therefore the improved LZW algorithm is suitable method in compressing Thai text.

ศูนย์วิทยทรัพยากร  
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา วิศวกรรมคอมพิวเตอร์ 10105  
สาขาวิชา วิทยาการคอมพิวเตอร์ 10105  
ปีการศึกษา 2533

ลายมือชื่อนิสิต ระวี วัฒนวิเศษ  
ลายมือชื่ออาจารย์ที่ปรึกษา [Signature]  
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม [Signature]



กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดีของ อาจารย์  
จารุมาตร ปันทอง อาจารย์ที่ปรึกษาวิทยานิพนธ์และ ผศ.ดร. วีระ ธีรวิทักษ์ อาจารย์ที่ปรึกษา  
วิทยานิพนธ์ร่วม จึงขอขอบพระคุณอาจารย์ทั้งสองมา ณ ที่นี้

ขอขอบพระคุณ รศ. ยืน ภู่วรวรรณ ที่ได้อนุเคราะห์ข้อมูลสำหรับการทำวิทยานิพนธ์  
ขอขอบพระคุณสำนักคอมพิวเตอร์ มหาวิทยาลัยมหิดล ที่ได้เอื้อเฟื้ออุปกรณ์ปริ้นท์ ขอขอบพระคุณ  
บัณฑิตวิทยาลัย ที่ให้ทุนอุดหนุนการวิจัย ขอขอบคุณเพื่อน ๆ ที่ให้ความช่วยเหลือเป็นอย่างดี  
ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณบิดามารดาที่ได้ให้กำลังใจแก่ผู้วิจัยเสมอมา

เรวดี ลิ้มปิไชติกุล


ศูนย์วิทยทรัพยากร  
จุฬาลงกรณ์มหาวิทยาลัย



## สารบัญ

	หน้า
บทคัดย่อภาษาไทย .....	ง
บทคัดย่อภาษาอังกฤษ .....	จ
กิตติกรรมประกาศ .....	ฉ
สารบัญตาราง .....	ณ
สารบัญรูป .....	ญ
บทที่	
1. บทนำ	
การลดขนาดข้อมูล .....	1
ลำดับชั้นของวิธีการอัดข้อมูล .....	3
แนวทางการวิจัย .....	6
ขอบเขตการวิจัย .....	7
ขั้นตอนการดำเนินการวิจัย .....	7
ประโยชน์ที่ได้รับจากการวิจัย .....	7
2. การอัดข้อมูล	
ทฤษฎีพื้นฐานของการอัดข้อมูล .....	8
อัลกอริทึมการอัดข้อมูลที่ไม่ขึ้นกับที่แมนติค .....	16
อัลกอริทึมที่มีการนำไปใช้ในปัจจุบัน .....	26
3. การอัดข้อความภาษาไทย	
ความเหลือเฟือในข้อความภาษาไทย .....	28
การอัดข้อความภาษาไทยที่ขึ้นกับที่แมนติค .....	29
การอัดข้อความภาษาไทยที่ไม่ขึ้นกับที่แมนติค .....	30
แนวทางการวิจัยการอัดข้อความภาษาไทย .....	30
4. การปรับเปลี่ยนการอัดข้อความภาษาไทย	
การปรับเปลี่ยนอัลกอริทึมฮัฟแมน .....	31
การปรับเปลี่ยนอัลกอริทึมแอลแซดดับบิว .....	37
การวิเคราะห์เชิงประสิทธิภาพของอัลกอริทึมฮัฟแมนที่ปรับเปลี่ยน .....	42
การวิเคราะห์เชิงประสิทธิภาพของอัลกอริทึมแอลแซดดับบิว ที่ปรับเปลี่ยน .....	42

5. การทดสอบประสิทธิภาพการอัดข้อมูล	
ข้อมูลที่ใช้ในการทดสอบ .....	44
ขั้นตอนการทดสอบ .....	44
การทดสอบข้อความภาษาไทยล้วน .....	45
การทดสอบข้อความภาษาไทยทั่ว ๆ ไป .....	45
6. สรุปผลการวิจัยและข้อเสนอแนะ	
สรุปผลการวิจัย .....	61
ข้อเสนอแนะ .....	63
บรรณานุกรม .....	64
ภาคผนวก .....	67
ประวัติผู้เขียน .....	82


  
 ศูนย์วิทยุทรัพยากร  
 จุฬาลงกรณ์มหาวิทยาลัย



สารบัญตาราง

	หน้า
ตารางที่ 5.1 การทดสอบประสิทธิภาพการอัดข้อมูลภาษาไทยล้วน เปรียบเทียบระหว่างวิธีการเดิมและวิธีการที่ปรับเปลี่ยน ที่ใช้คำไทย 255 คำ .....	47
ตารางที่ 5.2 การทดสอบเวลาที่ใช้ในการอัดและการขยายข้อมูลภาษาไทยล้วน เปรียบเทียบระหว่างวิธีการเดิมและวิธีการที่ปรับเปลี่ยน ที่ใช้คำไทย 255 คำ .....	49
ตารางที่ 5.3 การทดสอบประสิทธิภาพการอัดข้อมูลภาษาไทยล้วน เปรียบเทียบระหว่างวิธีที่ปรับเปลี่ยนที่ใช้คำไทย 255 คำ และ 511 คำ .....	52
ตารางที่ 5.4 การทดสอบเวลาที่ใช้ในการอัดและการขยายข้อมูลภาษาไทยล้วน เปรียบเทียบระหว่างวิธีที่ปรับเปลี่ยนที่ใช้คำไทย 255 คำ และ 511 คำ .....	53
ตารางที่ 5.5 การทดสอบประสิทธิภาพการอัดข้อมูลภาษาไทยทั่ว ๆ ไป เปรียบเทียบระหว่างวิธีการเดิมและวิธีการที่ปรับเปลี่ยน ที่ใช้คำไทย 255 คำ .....	54
ตารางที่ 5.6 การทดสอบเวลาที่ใช้ในการอัดและการขยายข้อมูลภาษาไทยทั่ว ๆ ไป เปรียบเทียบระหว่างวิธีการเดิมและวิธีการที่ปรับเปลี่ยน ที่ใช้คำไทย 255 คำ .....	56
ตารางที่ 5.7 การทดสอบประสิทธิภาพการอัดข้อมูลภาษาไทยทั่ว ๆ ไป เปรียบเทียบระหว่างวิธีที่ปรับเปลี่ยนที่ใช้คำไทย 255 คำ และ 511 คำ .....	59
ตารางที่ 5.8 การทดสอบเวลาที่ใช้ในการอัดและการขยายข้อมูลภาษาไทยทั่ว ๆ ไป เปรียบเทียบระหว่างวิธีที่ปรับเปลี่ยนที่ใช้คำไทย 255 คำ และ 511 คำ .....	60

## สารบัญรูป

			หน้า
รูปที่ 1.1	วิธีการลดขนาดข้อมูล .....		2
รูปที่ 1.2	วิธีการอัดข้อมูลที่ไม่ขึ้นกับซีแมนติค .....		5
รูปที่ 2.1	ตัวอย่างการแทนรหัส .....		9
รูปที่ 2.2	อัลกอริทึมการอัดข้อมูล .....		10
รูปที่ 2.3	ฟังก์ชันของสารสนเทศ .....		12
รูปที่ 2.4	วิธีการของชานนอน-ฟาโน .....		17
รูปที่ 2.5	วิธีการของฮัฟแมน .....		18
รูปที่ 2.6	รูปแบบความน่าจะเป็นของข้อมูล .....		20
รูปที่ 2.7	การเข้ารหัสและการสร้างลิสต์โดยอัลกอริทึมบีเอสทีดับบิว .....		24
รูปที่ 2.8	การย้ายไปหน้าของรหัส a ในลิสต์ .....		24
รูปที่ 2.9	การสร้างรหัสใหม่ที่ได้จากอ่านข้อมูล ababccabcc .....		25
รูปที่ 2.10	การเข้ารหัสข้อมูล .....		26
รูปที่ 4.1	การอัดข้อมูลโดยอัลกอริทึมฮัฟแมน .....		31
รูปที่ 4.2	รูปแบบของรหัสนำหน้า .....		32
รูปที่ 4.3	การอัดข้อมูลโดยอัลกอริทึมฮัฟแมนที่ปรับเปลี่ยน .....		33
รูปที่ 4.4	ตารางรหัส ตารางค่าไทย และฮัฟแมนทรี .....		34
รูปที่ 4.5	ข้อมูลที่เข้ารหัส .....		34
รูปที่ 4.6	การขยายข้อมูลโดยอัลกอริทึมฮัฟแมนที่ปรับเปลี่ยน .....		35
รูปที่ 4.7	ฟังก์ชันแฮช .....		38
รูปที่ 4.8	ตารางที่ใช้ในการเข้ารหัส .....		38
รูปที่ 4.9	ตารางที่ใช้ในการถอดรหัส .....		38
รูปที่ 4.10	การกระจายค่าไทย .....		39
รูปที่ 4.11	ตารางรหัสของอัลกอริทึมแอลแซดดับบิวที่ปรับเปลี่ยน .....		39
รูปที่ 4.12	ตารางรหัสและข้อมูลที่เข้ารหัสโดยอัลกอริทึมแอลแซดดับบิว .....		40
รูปที่ 4.13	ตารางรหัสและข้อมูลที่เข้ารหัสโดยอัลกอริทึมแอลแซดดับบิว ที่ปรับเปลี่ยน .....		41
รูปที่ 5.1	กราฟแสดงประสิทธิภาพการอัดข้อมูลภาษาไทยล้วน เปรียบเทียบระหว่างวิธีการเดิมและวิธีการที่ปรับเปลี่ยน ที่ใช้ค่าไทย 255 คำ .....		48

รูปที่ 5.2	กราฟแสดงเวลาที่ใช้ในการอัดข้อมูลภาษาไทยล้วน เปรียบเทียบระหว่างวิธีการเดิมและวิธีการที่ปรับเปลี่ยน ที่ใช้คำไทย 255 คำ .....	50
รูปที่ 5.3	กราฟแสดงเวลาที่ใช้ในการขยายข้อมูลภาษาไทยล้วน เปรียบเทียบระหว่างวิธีการเดิมและวิธีการที่ปรับเปลี่ยน ที่ใช้คำไทย 255 คำ .....	51
รูปที่ 5.4	กราฟแสดงประสิทธิภาพการอัดข้อมูลภาษาไทยทั่ว ๆ ไป เปรียบเทียบระหว่างวิธีการเดิมและวิธีการที่ปรับเปลี่ยน ที่ใช้คำไทย 255 คำ .....	55
รูปที่ 5.5	กราฟแสดงเวลาที่ใช้ในการอัดข้อมูลภาษาไทยทั่ว ๆ ไป เปรียบเทียบระหว่างวิธีการเดิมและวิธีการที่ปรับเปลี่ยน ที่ใช้คำไทย 255 คำ .....	57
รูปที่ 5.6	กราฟแสดงเวลาที่ใช้ในการขยายข้อมูลภาษาไทยทั่ว ๆ ไป เปรียบเทียบระหว่างวิธีการเดิมและวิธีการที่ปรับเปลี่ยน ที่ใช้คำไทย 255 คำ .....	58