

การประยุกต์ใช้การโปรแกรมตรรกะเชิงอุปนัยในการรู้จำตัวพิมพ์อักษรภาษาไทย

นางสาว อภิญญา สุพรรณวรรณษา



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาคตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2540

ISBN 974-637-943-7

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

**AN APPLICATION OF INDUCTIVE LOGIC PROGRAMMING
TO THAI PRINTED CHARACTER RECOGNITION**

Miss Apinya Supanwassa

**A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science**

Department of Computer Engineering

Graduate School

Chulalongkorn University

Academic Year 1997

ISBN 974-637-943-7

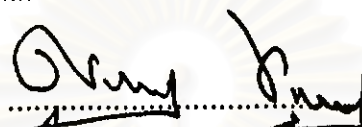
หัวข้อวิทยานิพนธ์ การประยุกต์ใช้การโปรแกรมตรรกะเชิงอุปนัยในการรู้จำตัวพิมพ์อักษรภาษาไทย

โดย นางสาว อภิญญา สุพรรณวรรณมา


ภาควิชา วิศวกรรมคอมพิวเตอร์

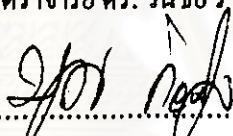
อาจารย์ที่ปรึกษา อาจารย์ ดร. บุญเสริม กิจศิริกุล

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้บัณฑิตวิทยาลัยฉบับนี้เป็นส่วนหนึ่งของการ
ศึกษาตามหลักสูตรปริญญามหาบัณฑิต



..... คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ นายแพทย์ สุภวัฒน์ ชูติวงศ์)

คณะกรรมการสอบวิทยานิพนธ์


..... ประธานกรรมการ
(รองศาสตราจารย์ ดร. วันชัย รวีไพบูลย์)


..... อาจารย์ที่ปรึกษา
(อาจารย์ ดร. บุญเสริม กิจศิริกุล)


..... กรรมการ
(อาจารย์ ดร. สืบสกุล พิภพมงคล)


..... กรรมการ
(ผู้ช่วยศาสตราจารย์ นงลักษณ์ ไคววารัช)

สถาบันวิจัยปฏิบัติการ
จุฬาลงกรณ์มหาวิทยาลัย

อภิญา สุพรรณวรรณา : การประยุกต์ใช้การโปรแกรมตรรกะเชิงอุปนัยในการรู้จำตัวพิมพ์อักษรภาษาไทย
(AN APPLICATION OF INDUCTIVE LOGIC PROGRAMMING TO THAI PRINTED CHARACTER
RECOGNITION) อ.ที่ปรึกษา : อาจารย์ ดร. บุญเสริม กิจศิริกุล , 78 หน้า. ISBN 974-637-943-7.

งานวิทยานิพนธ์ฉบับนี้มีวัตถุประสงค์ เพื่อประยุกต์ใช้การโปรแกรมตรรกะเชิงอุปนัย หรือ ไอแอลพี ในการรู้จำตัวพิมพ์อักษรภาษาไทย ระบบไอแอลพีที่เลือกใช้คือ ระบบ PROGOL ขั้นตอนการวิจัยเริ่มจากการใช้ PROGOL ในการเรียนรู้ตัวพิมพ์อักษรภาษาไทย ซึ่งข้อมูลที่เป็นต่อการเรียนรู้ ได้แก่ ตัวอย่าง และ ความรู้ส่วนหลัง ผลที่ได้จากการเรียนรู้คือ กลุ่มของกฎ ซึ่งแต่ละกฎจะนิยามลักษณะสำคัญของตัวอักษรภาษาไทยแต่ละตัว ขั้นตอนถัดมา คือ การนำกฎที่ได้จากการเรียนรู้มาใช้ในการรู้จำตัวอักษร โดยเปรียบเทียบกฎที่ได้กับตัวอักษรที่ต้องการรู้จำ และเลือกกฎที่ตรงกับลักษณะของตัวอักษรนั้นๆมากที่สุดให้เป็นผลการรู้จำ การทดสอบวิธีการดังกล่าวแบ่งออกเป็น 2 การทดลอง การทดลองแรกเพื่อทดสอบการรู้จำตัวอักษรในรูปแบบที่ไม่เคยเรียนรู้มาก่อน โดยใช้ตัวอักษรรูปแบบ EUCROSIA ในการเรียนรู้ และ ใช้รูปแบบ CORDIA ในการทดสอบการรู้จำ พบว่าผลการรู้จำมีความถูกต้อง 87.38 % จากจำนวนตัวอักษรที่ทำการทดสอบ 539 ตัวอักษร การทดลองที่สองเพื่อทดสอบการรู้จำตัวอักษรที่มีสัญญาณรบกวน โดยใช้ตัวอักษรรูปแบบ CORDIA และ EUCROSIA ในการเรียนรู้ และ นำตัวอักษรทั้งสองรูปแบบนั้นไปคัดลอกด้วยเครื่องถ่ายเอกสารได้จำนวนตัวอักษร 2156 ตัวอักษร แล้วจึงนำมาทดสอบการรู้จำ พบว่าผลการรู้จำมีความถูกต้อง 87.89 % เวลาในการรู้จำโดยเฉลี่ย 0.13 วินาทีต่อการรู้จำ 1 ตัวอักษร

สถาบันวิทยบริการ

จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา
สาขาวิชา
ปีการศึกษา

วิศวกรรมคอมพิวเตอร์
วิทยาศาสตร์คอมพิวเตอร์
2540

ลายมือชื่อนิติกร อภิญา สุพรรณวรรณา
ลายมือชื่ออาจารย์ที่ปรึกษา
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

C718241 : MAJOR COMPUTER SCIENCE
KEY WORD:

INDUCTIVE LOGIC PROGRAMMING / MACHINE LEARNING / PROGOL / RECOGNITION / THAI CHARACTERS

APINYA SUPANWASSA : AN APPLICATION OF INDUCTIVE LOGIC PROGRAMMING TO THAI PRINTED CHARACTER RECOGNITION. THESIS ADVISOR : DR. BOONSERM KIJSIRIKUL.

78 pp. ISBN 974-637-943-7.

The purpose of this thesis is to apply Inductive Logic Programming (ILP) to Thai printed character recognition. The ILP system which have been chosen is PROGOL. First, PROGOL is employed to learn Thai printed characters. The examples of characters and background knowledge are mainly used to train PROGOL. The output of PROGOL is a set of rules each of which defines the characteristics of a Thai character. Then, the learned rules are used to recognize an input character by comparing them with the input character, and the most matched rule together with associated character is selected as the output. Two experiments were run to test the method. In the first experiment designed for unseen fonts, the Eucrosia fonts are used for training and the Cordia fonts are used for testing. The recognition rate is 87.38%, tested with 539 characters. In the second experiment for noisy fonts, the Cordia and Eucrosia fonts are used for training, and the copies produced by a copy machine of both types of fonts composed of 2156 characters are used for testing. The recognition rate is 87.89%. The average recognition time is 0.13 second per character.



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา.....วิศวกรรมคอมพิวเตอร์
สาขาวิชา.....วิทยาศาสตร์คอมพิวเตอร์
ปีการศึกษา.....2540

ลายมือชื่อนิสิต..... อภิญญา สุทธิธรรม
ลายมือชื่ออาจารย์ที่ปรึกษา.....
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ได้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของ อาจารย์ ดร.บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้คำแนะนำและข้อคิดเห็นต่างๆ ของการวิจัยมาด้วยดี ตลอด ขอขอบคุณ รองศาสตราจารย์ ดร.สมชาย จิตะพันธ์กุล และ คุณเคชา รัตนธาร ที่อนุญาตให้ใช้ โปรแกรมการประมวลผลขั้นต้นจากงานวิทยานิพนธ์ของคุณเคชา รัตนธาร ได้ ขอขอบคุณ คุณอดิศร ทิพย์ไพศาล ที่ให้ใช้เครื่องกวาดตรวจในการเก็บข้อมูลภาพด้วยอักษร รวมทั้งคุณสรรชัย ธนาชัยแสง และ เพื่อนๆที่คอยให้กำลังใจมาโดยตลอด

ทำนนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา-มารดา ซึ่งได้ให้การสนับสนุน และให้กำลังใจแก่ผู้ วิจัยเสมอมาจนสำเร็จการศึกษา



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญ

	หน้า
บทคัดย่อภาษาไทย.....	ง
บทคัดย่อภาษาอังกฤษ.....	จ
กิตติกรรมประกาศ.....	ค
สารบัญตาราง.....	ฅ
สารบัญภาพ.....	ญ
บทที่	
1. บทนำ.....	1
ความเป็นมาของงาน.....	1
วัตถุประสงค์.....	2
ขอบเขตของการวิจัย.....	2
ขั้นตอนการวิจัย.....	3
ประโยชน์ที่จะได้จากการวิจัย.....	3
งานวิจัยที่เกี่ยวข้อง.....	3
1. งานวิจัยด้าน ไอแอลที.....	3
2. งานวิจัยทางด้านการรู้จำตัวพิมพ์อักษรภาษาไทย.....	4
2. ทฤษฎีที่ใช้ในการวิจัย.....	9
การเรียนรู้ของเครื่อง.....	9
การโปรแกรมตรรกะเชิงอุปนัย.....	10
ระบบ PROGOL.....	11
1. ความหมายของระบบ PROGOL.....	11
2. การใช้งานระบบ PROGOL.....	11
3. ผลลัพธ์ที่ได้จากระบบ PROGOL.....	14
4. ขั้นตอนการเงินเนอรัล ไรซ้อนุประโยคของ PROGOL.....	15
การรู้จำแบบ.....	15
3. กรรมวิธีการเรียนรู้ตัวพิมพ์อักษรภาษาไทย.....	17
โครงสร้างของระบบ.....	17
1. การปรับปรุงคุณภาพของข้อมูล.....	18
1.1 การกำจัดสัญญาณรบกวน.....	19

สารบัญ (ต่อ)

บทที่	หน้า
1.2 การทำตัวอักษรให้บาง.....	19
1.3 การปรับกรอบของตัวอักษร.....	20
2. การวิเคราะห์หลักขณะสำคัญของตัวอักษร.....	20
2.1 การเข้ารหัสจุดภาพของตัวอักษร.....	20
2.2 การแปลงจุดภาพที่เข้ารหัสให้เป็นเวกเตอร์.....	24
2.3 การเปลี่ยนเวกเตอร์ให้เป็นหน่วยสร้างพื้นฐาน.....	25
2.4 การแบ่งระดับ และแบ่งเขตย่อยของตัวอักษร.....	26
2.5 การหาส่วนหัวของตัวอักษร.....	27
3. การเรียนรู้โดยใช้ระบบไอแอลที.....	28
3.1 ข้อมูลที่ใช้ในการเรียนรู้.....	28
3.2 ความรู้ส่วนหลัง.....	31
3.3 ข้อมูลตัวอย่าง.....	33
3.4 การกำหนดรูปแบบของสัญลักษณ์.....	35
ผลการเรียนรู้.....	36
4. การทดสอบการรู้จำตัวพิมพ์อักษรภาษาไทย.....	49
วิธีการทดสอบ.....	49
ผลการทดสอบ.....	50
1. การรู้จำตัวอักษรในรูปแบบที่ไม่เคยผ่านการเรียนรู้มาก่อน.....	50
2. การรู้จำตัวอักษรที่มีสัญญาณรบกวนปะปน.....	52
ปัญหาและข้อจำกัด.....	55
5. สรุปการวิจัย และข้อเสนอแนะ.....	57
สรุปการวิจัย.....	57
ข้อเสนอแนะ.....	57
รายการอ้างอิง.....	59
ภาคผนวก ก. ตัวอักษรภาษาไทยที่ใช้ในการวิจัย.....	61
ภาคผนวก ข. การกำหนดรูปแบบของสัญลักษณ์.....	62
ภาคผนวก ค. การกำหนดความรู้ส่วนหลัง.....	64
ภาคผนวก ง. เพิ่มเติมรายละเอียดการใช้งานในระบบ PROGOL.....	75
ประวัติผู้เขียน.....	78

สารบัญตาราง

ตารางที่	หน้า
3.1 ก. สูตรคำนวณหาจุดตัดไปของจุดต่อเนื่อง.....	22
ข. สูตรคำนวณหาจุดตัดไปของจุดปลาย.....	22
3.2 มุมของรหัสเงื่อนไข.....	23
3.3 มุมที่ได้จากการเรียงรหัสเงื่อนไข.....	23
3.4 รหัสต่อเนื่อง และ รหัสไม่ต่อเนื่อง ของรหัสเงื่อนไข.....	24
3.5 ผลที่ได้จากการเรียนรู้ตัวพิมพ์อักษรภาษาไทย รูปแบบ EUCROSIA ขนาด 48, 36, 32, 28, 24, 22 และ 20.....	37
3.6 ผลที่ได้จากการเรียนรู้ตัวพิมพ์อักษรภาษาไทย รูปแบบ CORDIA และ EUCROSIA ขนาด 48, 36, 32, 28, 24, 22 และ 20.....	42
4.1 ผลการทดสอบการรู้จำตัวอักษรในรูปแบบที่ไม่เคยผ่านการเรียนรู้มาก่อน.....	51
4.2 ผลการทดสอบการรู้จำตัวอักษรที่มีสัญญาณรบกวนปะปน.....	53

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญภาพ

รูปที่	หน้า
3.1 โครงสร้างของระบบการเรียนรู้.....	17
3.2 โครงสร้างของระบบการรู้จำ.....	18
3.3 การกำจัดสัญญาณรบกวน.....	19
ก. ก่อนการกำจัดสัญญาณรบกวน.....	19
ข. หลังการกำจัดสัญญาณรบกวน.....	19
3.4 การทำตัวอักษรให้บาง.....	19
ก. ตัวอักษรก่อนการทำให้บาง.....	19
ข. ตัวอักษรหลังการทำให้บาง.....	19
3.5 การปรับกรอบของตัวอักษร.....	20
ก. ทิศทางการตรวจสอบกรอบของตัวอักษร.....	20
ข. การปรับกรอบของตัวอักษรให้พอดีกับตัวอักษร.....	20
3.6 การแทนค่าจุดภาพด้วยรหัสเงื่อนไขที่เป็นเลขลบ ให้กับจุดภาพที่เป็นจุดปลาย.....	21
3.7 การแทนค่าจุดภาพด้วยรหัสเงื่อนไขที่เป็นเลขบวก ให้กับจุดภาพที่เป็นจุดต่อเนื่อง.....	21
3.8 การแปลงจากเวกเตอร์เส้นตรงให้เป็นเวกเตอร์วงกลม.....	24
ก. ก่อนทำให้เป็นวงกลม.....	24
ข. หลังการทำให้เป็นวงกลม.....	24
3.9 โครงสร้างของต้นไม้.....	25
ก. โครงสร้างของต้นไม้.....	25
ข. โครงสร้างของโหนด.....	25
3.10 การแปลงข้อมูลภาพที่เข้ารหัสเงื่อนไขให้เป็นเวกเตอร์.....	25
ก. ข้อมูลภาพที่เข้ารหัสเงื่อนไข.....	25
ข. เวกเตอร์ของตัวอักษร.....	25
3.11 หน่วยสร้างพื้นฐาน.....	26
ก. หน่วยสร้างพื้นฐานเส้นตรง.....	26
ข. หน่วยสร้างพื้นฐานวงกลม.....	26
3.12 เขตย่อยของตัวอักษร.....	27
3.13 ลำดับความสำคัญของเขตย่อยในการหาหัวของตัวอักษร.....	28
ก. จุดปลายที่เป็นเวกเตอร์วงกลม.....	28
ข. จุดปลายที่เป็นเวกเตอร์เส้นตรง.....	28
3.14 การแทนค่าข้อมูลส่วนย่อยของตัวอักษร.....	29

สารบัญญากาศ (ต่อ)

รูปที่	หน้า
3.15 ลักษณะส่วนหยักลง และ ส่วนหยักขึ้น ของตัวอักษร.....	30
ก. ส่วนหยักลง.....	30
ข. ส่วนหยักขึ้น.....	30
4.1 กราฟเปรียบเทียบเปอร์เซ็นต์ความถูกต้อง การรู้จำตัวอักษรรูปแบบที่ไม่เคยเรียนรู้มาก่อน.....	51
4.2 กราฟเปรียบเทียบเวลาที่ใช้ในการรู้จำตัวอักษรในรูปแบบที่ไม่เคยเรียนรู้มาก่อน.....	52
4.3 กราฟเปรียบเทียบเปอร์เซ็นต์ความถูกต้อง การรู้จำตัวอักษรที่มีสัญญาณรบกวนปะปน.....	54
4.4 กราฟเปรียบเทียบเวลาที่ใช้ในการรู้จำตัวอักษรที่มีสัญญาณรบกวนปะปน.....	55



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย