



242480

การเปรียบเทียบทกนิคการลดน้ำหนักการจำแนกเอกสารบน  
โครงข่ายประสาท

ศุภชัย มอดคาสันกุ

วิทยาศาสตรมหาบัณฑิต  
สาขาวิชาภาษาคอมพิวเตอร์

บัณฑิตวิทยาลัย  
มหาวิทยาลัยเชียงใหม่  
กุมภาพันธ์ 2552



242480

การเปรียบเทียบทكنิคการลดมิติสำหรับการจำแนกเอกสารบน  
โครงข่ายประสาท

ศุภชัย มุกดานนิท



วิทยานิพนธ์นี้เสนอต่อนบณฑิตวิทยาลัยเพื่อเป็นส่วนหนึ่ง  
ของการศึกษาตามหลักสูตรปริญญา  
วิทยาศาสตรมหาบัณฑิต  
สาขาวิชาภาษาการคอมพิวเตอร์

บัณฑิตวิทยาลัย  
มหาวิทยาลัยเชียงใหม่  
ถุมภาพันธ์ 2552

## การเปรียบเทียบทekenikการลดมิติสำหรับการจำแนกเอกสารบนโครงข่ายประสาท

ศุภชัย มุกดาสนิท

วิทยานิพนธ์ได้รับการพิจารณาอนุมัติให้นับเป็นส่วนหนึ่งของการศึกษา  
ตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต  
สาขาวิชาวิทยาการคอมพิวเตอร์

คณะกรรมการสอบวิทยานิพนธ์

ประธานกรรมการ

ผู้ช่วยศาสตราจารย์ ดร. จริรยุทธ ไชยจารุวนิช

กรรมการ

ผู้ช่วยศาสตราจารย์ ดร. เสมอแข สมหอม

กรรมการ

รองศาสตราจารย์ ดร. กฤษณะ ไวยมัย

3 กุมภาพันธ์ 2552

© ลิขสิทธิ์ของมหาวิทยาลัยเชียงใหม่

## กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลงได้ด้วยความกรุณาจาก ผู้ช่วยศาสตราจารย์ ดร.เสมอแเขต สมหอน อาจารย์ที่ปรึกษาวิทยานิพนธ์ และผู้ช่วยศาสตราจารย์ ดร.จิรยุทธ ไชยจารุวนิช ผู้ชี้งกรุณากล่าวให้ความรู้ คำปรึกษา คำแนะนำ และตรวจสอบแก้ไขจนวิทยานิพนธ์เสร็จสมบูรณ์ ผู้เขียนขอกราบขอบพระคุณ เป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.จิรยุทธ ไชยจารุวนิช และ รองศาสตราจารย์ ดร.กฤษณะ ไวยนัย ที่กรุณารับเป็นคณะกรรมการสอบวิทยานิพนธ์ และ ให้คำแนะนำเป็นอย่างดี

ขอบคุณภาควิชาจิตวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ที่อนุเคราะห์ให้ใช้อุปกรณ์คอมพิวเตอร์และสถานที่ในการประกอบการศึกษาด้านกว้าง ขอขอบคุณ เจ้าหน้าที่ทุกคนที่อำนวยความสะดวกให้ทำงานได้โดยราบรื่น

ขอขอบคุณบัณฑิตวิทยาลัย มหาวิทยาลัยเชียงใหม่ ที่ให้การสนับสนุนค่าใช้จ่ายในการทำวิทยานิพนธ์

ขอขอบคุณคณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ที่ให้การสนับสนุนทุนการศึกษาและค่าใช้จ่ายในการเสนอผลงานทางวิชาการ

ขอกราบขอบพระคุณ คุณพ่อ คุณแม่ ที่ได้อบรมสั่งสอน และคณาจารย์ทุกท่านที่ช่วยประสิทธิ์ประสาทวิชาความรู้ให้ผู้เขียนตั้งแต่ระดับอนุบาลจนกระทั่งถึงปัจจุบัน

ศุภชัย นุกคำสนิท

ชื่อเรื่องวิทยานิพนธ์	การเปรียบเทียบเทคนิคการลดมิติสำหรับการจำแนก เอกสารบนโครงข่ายภาษาไทย
ผู้เขียน	นายศุภชัย นุกดาสนิท
ปริญญา	วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผู้ช่วยศาสตราจารย์ ดร. เสนอแนะ สมหอน

242480

บทคัดย่อ

การจำแนกเอกสารมีความสำคัญเนื่องจากจำนวนเอกสารที่มากมายและ การจัดการเอกสาร ที่ยุ่งยาก งานวิจัยนี้จึงได้เสนอกระบวนการ การจำแนกเอกสาร โดยใช้โครงข่ายภาษาไทย ซึ่งเป็นที่ ทราบกันดีว่า โครงข่ายภาษาไทยมีความสามารถสูงในการจำแนกข้อมูลในงานด้านต่างๆรวมถึงการ จำแนกเอกสารด้วย เมื่อว่า โครงข่ายภาษาไทยมีความสามารถในการเรียนรู้และจัดการความซับซ้อน แต่การลดความซับซ้อนของการคำนวณของ โครงข่ายภาษาไทย โดยการจัดการกับจำนวนของคำ สำคัญในเอกสาร ที่มีปริมาณมากก็ยังมีความจำเป็นซึ่งจะส่งผลถึงประสิทธิภาพของการจำแนก เอกสาร งานวิจัยนี้มีการนำเทคนิคการลดมิติสามเหลี่ยมtechnique ประกอบด้วยพีแอลเอส ด้านความหมาย ภาษาใน และรีเดเวนซ์สกอร์ มาใช้ในการศึกษาเปรียบเทียบและวัดประสิทธิภาพในการใช้การลดมิติ แบบต่างๆกับการจำแนกเอกสาร ข้อมูลตัวอย่างที่ใช้ในการสร้างตัวจำแนกและใช้ทดสอบ ประสิทธิภาพคือเอกสารชีวสารสนเทศ ผลที่ได้คือ การลดมิติด้วยวิธีด้านความหมายภาษาใน และ พีแอลเอสส่งผลให้การจำแนกเอกสารมีความถูกต้องที่ใกล้เคียงกัน ซึ่งสูงกว่าวิธีรีเดเวนซ์สกอร์ ด้านเวลาการเรียนรู้พีแอลเอส สามารถลดเวลาการเรียนรู้ของ โครงข่ายภาษาไทยได้ที่สุดรองลงมา คือรีเดเวนซ์สกอร์ และสุดท้ายคือรีเดเวนซ์สกอร์

**Thesis Title** A Comparison of Dimensionality Reduction Techniques  
for Document Classification on Neural Network

**Author** Mr. Supachai Mukdasanit

**Degree** Master of Science (Computer Science)

**Thesis Advisor** Assistant Professor Dr. Samerkae Somhom

**ABSTRACT**

**242480**

The document classification is a necessity due to the very large amount of documents and the complexity management of documents. This research presents the document classification process using neural network. The neural network has been recognized as one of the most successful classification methods for many applications including document classification. Even though the learning ability and computational complexity of training in neural network, reducing computational complexity is an essential issue to efficiently handle a large number of terms in practical applications of document classification. This research provides a comparison study of three dimension reduction techniques, namely Partial Least Squares (PLS), Latent Semantic Indexing (LSI) and Relevancy Score (RE), and evaluates the relative performance of classification procedures incorporating those methods. The sample documents are the bio-information documents to create classifier and evaluate performance. The results show that LSI and PLS can improve accuracy higher than RE for classifying by neural network. In the training time, the PLS can improve the training time better of NN than LSI and RE respectively.

## สารบัญ

	หน้า
<b>กิตติกรรมประกาศ</b>	ค
<b>บทคัดย่อภาษาไทย</b>	๔
<b>บทคัดย่อภาษาอังกฤษ</b>	๕
<b>สารบัญตาราง</b>	๗
<b>สารบัญภาพ</b>	๘
<b>บทที่ 1 บทนำ</b>	๑
1.1 หลักการและเหตุผล	๑
1.2 วัตถุประสงค์ของการวิจัย	๒
1.3 ประโยชน์ที่ได้รับจากการศึกษา เชิงทฤษฎี และเชิงประยุกต์	๒
1.4 ขอบเขตการวิจัย	๒
1.5 วิธีการวิจัย	๓
1.6 อุปกรณ์ที่ใช้ในการวิจัย	๓
1.7 สถานที่ที่ใช้ในการดำเนินการวิจัยและรวบรวมข้อมูล	๓
<b>บทที่ 2 หลักการ และทฤษฎีที่เกี่ยวข้อง</b>	๔
2.1 การจำแนกเอกสาร	๔
2.2 การสร้างคัชณีเอกสาร	๔
2.3 การลดความมิติเอกสาร	๕
2.4 ทฤษฎีโครงข่ายประชากร	๑๔
<b>บทที่ 3 การทำดัชนีเอกสาร</b>	๑๙
3.1 หลักการการทำดัชนีเอกสาร	๑๙
3.2 การทดลองการทำดัชนีเอกสาร	๑๙
3.3 ผลการทดลองการทำดัชนีเอกสาร	๒๐
3.4 วิจารณ์และสรุปการทำดัชนีเอกสาร	๒๑

## สารบัญ (ต่อ)

	หน้า
<b>บทที่ 4 การลดนิติเอกสาร</b>	22
4.1 หลักการการลดนิติเอกสาร	22
4.2 การลดนิติเอกสารด้วยวิธีการเลือกลักษณะเด่นด้วยกระบวนการ รีແวเนช์สกอร์	22
4.3 การลดนิติด้วยวิธีการสกัดลักษณะเด่นด้วยวิธีแอลเอสไอ	23
4.4 การลดนิติด้วยวิธีการสกัดลักษณะเด่นด้วยวิธีพีแอลเอส	24
4.5 วิจารณ์และสรุปผล	27
<b>บทที่ 5 การจำแนกเอกสารด้วยแบบจำลองโครงข่ายประชาท</b>	29
5.1 หลักการการจำแนกเอกสารด้วยแบบจำลองโครงข่ายประชาท	29
5.2 การออกแบบการทดลอง	29
5.3 การวัดประสิทธิภาพ	33
5.4 ผลการทดลอง	33
5.5 วิจารณ์และสรุปผล	40
<b>บทที่ 6 สรุปผลการวิจัย</b>	42
6.1 บทสรุป	42
6.2 ปัญหาและอุปสรรค	44
6.3 ข้อเสนอแนะการทำงานวิจัย	44
<b>บรรณานุกรม</b>	45
<b>ประวัติผู้เขียน</b>	47

## สารบัญตาราง

ตาราง	หน้า
2.1 ตัวอย่างฟังก์ชันการคำนวณหาค่าน้ำหนักสำหรับการเลือกลักษณะเด่น	7
5.1 ผลการจำแนกเอกสารของ NN+term-doc	34
5.2 ผลการจำแนกเอกสารของ NN+PLS	34
5.3 ผลการจำแนกเอกสารของ NN+LSI	35
5.4 ผลการจำแนกเอกสารของ NN+RE	36
5.5 เวลาการเรียนรู้ข้อมูลของ NN+term-doc	37
5.6 เวลาการเรียนรู้ข้อมูลของ NN+PLS	38
5.7 เวลาการเรียนรู้ข้อมูลของ NN+LSI	38
5.8 เวลาการเรียนรู้ข้อมูลของ NN+RE	39

## สารบัญภาพ

หัวข้อ	หน้า
<b>รูป</b>	
2.1 แสดงการสร้างดัชนีเอกสาร	5
2.2 แสดงการหมุนแกน	10
2.3 แสดงการลดมิติ	10
2.4 แสดงรูปแบบความสัมพันธ์ของโครงสร้างภายในของกระบวนการพีเอลเอส	12
2.5 แสดงเซลล์ประสานของโครงข่ายประสาน	15
2.6 แสดงโครงข่ายประสานนิคการเรียนรู้แบบแพร์กัลับ	16
2.7 แสดงพื้นผิวของความผิดพลาด (Error surface) จากการปรับค่าน้ำหนักสองครั้ง	17
4.1 แสดงค่าความผิดพลาดเฉลี่ยคำถังสองของ $\delta_k$ มิติแรก	26
4.2 แสดงค่าความถูกต้องของการจำแนกที่สร้างตัวแทนเอกสาร 25 มิติ ในแต่ละรอบของการปรับค่า $\delta_k$ ในมิติที่ 1	27
5.1 แสดงเปอร์เซ็นต์ความถูกต้องของการจำแนกจากการใช้โนนคในชั้นกลางโดยมีข้อมูลเข้าคือเมตริกซ์คำสำคัญ-เอกสาร	31
5.2 แสดงเปอร์เซ็นต์ความถูกต้องของการจำแนกจากการใช้โนนคในชั้นกลางโดยมีข้อมูลเข้าคือตัวแทนเอกสารสร้างจากแอลเอสไอล 400 มิติ	31
5.3 แสดงเปอร์เซ็นต์ความถูกต้องของการจำแนกจากการใช้โนนคในชั้นกลางโดยมีข้อมูลเข้าคือตัวแทนเอกสารสร้างจากพีเอลเอส 25 มิติ	32
5.4 แสดงเปอร์เซ็นต์ความถูกต้องของการจำแนกจากการใช้โนนคในชั้นกลางโดยมีข้อมูลเข้าคือเมตริกซ์คำสำคัญ-เอกสาร ได้จากการเลือกคำสำคัญด้วยวิธีรีเลเวนช์สกอร์ 450 คำ	32
5.5 แสดงการเปรียบเทียบความถูกต้องสูงสุดของการจำแนกเอกสารแต่ละแบบ	40
5.6 แสดงการเปรียบเทียบเวลาการเรียนรู้ของการจำแนกเอกสารแต่ละแบบ	41