การเปรียบเทียบเทคนิคการลดมิติสำหรับการจำแนกเอกสารบนโครงข่ายประสาท

ศุภชัย มุกดาสนิท

วิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์

บัณฑิตวิทยาลัย
มหาวิทยาลัยเชียงใหม่
กุมภาพันธ์ 2552

การเปรียบเทียบเทคนิคการลดมิติสำหรับการจำแนกเอกสารบนโครงข่ายประสาท

ศุภชัย มุกดาสนิท

วิทยานิพนธ์นี้เสนอต่อบัณฑิตวิทยาลัยเพื่อเป็นส่วนหนึ่ง
ของการศึกษาตามหลักสูตรปริญญา
วิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์

บัณฑิตวิทยาลัย
มหาวิทยาลัยเชียงใหม่
กุมภาพันธ์ 2552

การเปรียบเทียบเทคนิคการลดมิติสำหรับการจำแนกเอกสารบนโครงข่ายประสาท

ศุภชัย มุกดาสนิท

วิทยานิพนธ์นี้ได้รับการพิจารณาอนุมัติให้นับเป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
ผู้ช่วยศาสตราจารย์ ดร.จีรยุทธ ไชยจารุวณิช

.....กรรมการ
ผู้ช่วยศาสตราจารย์ ดร.เสมอแข สมหอม

.....กรรมการ
รองศาสตราจารย์ ดร.กฤษณะ ไวยมัย

3 กุมภาพันธ์ 2552

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลงได้ด้วยความกรุณาจาก ผู้ช่วยศาสตราจารย์ ดร.เสมอแข สมหอม อาจารย์ที่ปรึกษาวิทยานิพนธ์ และผู้ช่วยศาสตราจารย์ ดร.จีรยุทธ ไชยจารุวณิช ผู้ซึ่งกรุณาให้ความรู้ คำปรึกษา คำแนะนำ และตรวจสอบแก้ไขจนวิทยานิพนธ์เสร็จสมบูรณ์ ผู้เขียนขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.จีรยุทธ ไชยจารุวณิช และ รองศาสตราจารย์ ดร.กฤษณะ ไวยมัย ที่กรุณารับเป็นคณะกรรมการสอบวิทยานิพนธ์ และ ให้คำแนะนำเป็นอย่างดี

ขอบคุณภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ที่อนุเคราะห์ให้ใช้อุปกรณ์คอมพิวเตอร์และสถานที่ ในการประกอบการศึกษาค้นคว้า ขอขอบคุณเจ้าหน้าที่ทุกคนที่อำนวยความสะดวกให้ทำงานได้โดยราบรื่น

ขอขอบคุณบัณฑิตวิทยาลัย มหาวิทยาลัยเชียงใหม่ ที่ให้การสนับสนุนค่าใช้จ่ายในการทำวิทยานิพนธ์

ขอขอบคุณคณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ที่ให้การสนับสนุนทุนการศึกษาและค่าใช้จ่ายในการเสนอผลงานทางวิชาการ

ขอกราบขอบพระคุณ คุณพ่อ คุณแม่ ที่ได้อบรมสั่งสอน และคณาจารย์ทุกท่านที่ช่วยประสิทธิ์ประสาทวิชาความรู้ให้ผู้เขียนตั้งแต่ระดับอนุบาลจนกระทั่งถึงปัจจุบัน

ศุภชัย มุกดาสนิท

ชื่อเรื่องวิทยานิพนธ์ การเปรียบเทียบเทคนิคการลดมิติสำหรับการจำแนกเอกสารบนโครงข่ายประสาท

ผู้เขียน นายศุภชัย มุกดาสนิท

ปริญญา วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.เสมอแข สมหอม

บทคัดย่อ

การจำแนกเอกสารมีความสำคัญเนื่องจากจำนวนเอกสารที่มากมายและ การจัดการเอกสารที่ยุ่งยาก งานวิจัยนี้จึงได้เสนอกระบวนการการจำแนกเอกสารโดยใช้โครงข่ายประสาท ซึ่งเป็นที่ทราบกันดีว่าโครงข่ายประสาทมีความสามารถสูงในการจำแนกข้อมูลในงานด้านต่างๆรวมถึงการจำแนกเอกสารด้วย แม้ว่าโครงข่ายประสาทมีความสามารถในการเรียนรู้และจัดการความซับซ้อน แต่การลดความซับซ้อนของการคำนวณของโครงข่ายประสาทโดยการจัดการกับจำนวนของคำสำคัญในเอกสาร ที่มีปริมาณมากก็ยังมีความจำเป็นซึ่งจะส่งผลถึงประสิทธิภาพของการจำแนกเอกสาร งานวิจัยนี้มีการนำเทคนิคการลดมิติสามเทคนิคประกอบด้วยพีแอลเอส ดัชนีความหมายภายใน และรีเลแวนซี่สกอร์ มาใช้ในการศึกษาเปรียบเทียบและวัดประสิทธิภาพในการใช้การลดมิติแบบต่างๆกับการจำแนกเอกสาร ข้อมูลตัวอย่างที่ใช้ในการสร้างตัวจำแนกและใช้ทดสอบประสิทธิภาพคือเอกสารชีวสารสนเทศ ผลที่ได้คือ การลดมิติด้วยวิธีดัชนีความหมายภายใน และพีแอลเอสส่งผลให้การจำแนกเอกสารมีความถูกต้องที่ใกล้เคียงกัน ซึ่งสูงกว่าวิธีรีเลแวนซี่สกอร์ ด้านเวลาการเรียนรู้พีแอลเอส สามารถลดเวลาการเรียนรู้ของโครงข่ายประสาทได้ดีที่สุดรองลงมาคือแอลเอสไอ และสุดท้ายคือรีเลแวนซี่สกอร์

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

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.

สารบาญ

สารบาญ (ต่อ)

สารบาญตาราง

สารบาญภาพ