บทที่ 1
บทนำ

ในบทนี้กล่าวถึง หลักการและเหตุผล วัตถุประสงค์งานวิจัย ประโยชน์ที่ได้รับและขอบเขตงานวิจัย โดยมีเนื้อหาดังต่อไปนี้

1.1 หลักการและเหตุผล

การเพิ่มขึ้นของเอกสารในงานด้านต่างๆ ทำให้เกิดปัญหาในการจัดการกับเอกสารที่เพิ่มขึ้น ซึ่งรวมไปถึงการจำแนกเอกสาร (Document classification) ให้เป็นหมวดหมู่ด้วยเช่นกัน ถ้าเราสามารถจำแนกเอกสารเป็นหมวดหมู่ ก็จะช่วยให้สามารถจัดการเอกสารได้อย่างมีประสิทธิภาพ เช่น การค้นคืนสารสนเทศ (Information retrieval) สามารถค้นคืนเอกสารได้ตรงกับความต้องการและรวดเร็ว แต่การจำแนกเอกสารทำได้ยากและใช้เวลานาน หากกระบวนการนี้กระทำโดยมนุษย์ เนื่องจากมีปัจจัยต่างๆ เข้ามาเกี่ยวข้อง เช่น ความเข้าใจในเนื้อหาของงานนั้นๆ และเวลาที่ใช้ในการอ่านเอกสาร ด้วยเหตุนี้จึงมีการนำเทคโนโลยีสารสนเทศมาประยุกต์ใช้กับการจำแนกเอกสาร

การจำแนกเอกสารมีวัตถุประสงค์คือ การปรับปรุงความถูกต้องและความรวดเร็วในการจำแนกเอกสาร อันเนื่องมาจาก จำนวนเอกสารที่มีมากมาย ประกอบกับจำนวนกลุ่มของเอกสารที่มีหลายกลุ่ม เมื่อแปลงเอกสารให้อยู่ในรูปของ เมตริกซ์ความถี่ของคำสำคัญ-เอกสาร ปัญหาที่เกิดขึ้นคือ จำนวนของคำสำคัญที่มีจำนวนมาก ทำให้ใช้เวลานานในการจำแนกเอกสาร และปัญหาเรื่องความหมายของคำเข้ามาเกี่ยวข้อง การแก้ปัญหาจำนวนคำสำคัญที่มีจำนวนมากแบ่งออกเป็นสองวิธี ประกอบด้วย วิธีเลือกลักษณะเด่น (Feature selection) คือวิธีเลือกคำสำคัญบางคำจากคำสำคัญทั้งหมด โดยพิจารณาจากค่าน้ำหนักของคำสำคัญนั้นๆ อีกวิธีคือ วิธีสกัดลักษณะเด่น (Feature extraction) เป็นวิธีการแปลงลักษณะของเวคเตอร์ความถี่ของคำสำคัญ ให้อยู่ในรูปแบบใหม่ที่มีจำนวนมิติที่น้อยลง ผลจากการสกัดลักษณะเด่นมีสองประการคือ สามารถลดจำนวนข้อมูลเข้า โดยมีน้อยกว่าคำสำคัญทั้งหมด และยังสามารถแก้ปัญหาคำที่มีความหมายเหมือนกัน (Synonym) หรือคำที่เขียนเหมือนกันแต่ความหมายแตกต่างกัน (Polysemy) ในระดับที่ยอมรับได้ (Barbara, 2000)

การแทนเอกสารให้อยู่ในรูปแบบที่มีขนาดเล็กลงด้วยการสกัดลักษณะเด่นที่เป็นที่นิยมและมีประสิทธิภาพสูง คือการหาโครงสร้างความหมายภายใน (Latent semantic structure) โดยการสร้างดัชนีความหมายภายใน (Latent semantic indexing) ซึ่งเป็นวิธีการหาความสัมพันธ์ของคำสำคัญภายในเอกสาร ด้วยเหตุนี้ จึงได้มีการนำดัชนีความหมายภายในมาสร้างตัวแทนเอกสาร โดย

อาศัยแบบจำลองในการจำแนก เช่น ซัพพอร์ทเวคเตอร์แมชชีน (Support vector machine) หรือแบบจำลองโครงข่ายประสาท (Neural network) แต่ดัชนีความหมายภายในใช้กระบวนการสร้างตัวแทนเอกสารด้วยเองโดยไม่มีการใช้ข้อมูลของกลุ่มเข้ามาเกี่ยวข้อง ดังนั้นกลุ่มข้อมูลที่มีขนาดเล็กจึงอาจไม่ถูกนำมาพิจารณาในการสร้างตัวแทนอกสาร ด้วยเหตุนี้ จึงได้มีการนำพีแอลเอส(Partial least square :PLS) มาใช้ในการสร้างตัวแทนเอกสาร ซึ่งการสร้างตัวแทนเอกสารนี้มีการใช้ข้อมูลของคำสำคัญของเอกสารและข้อมูลของกลุ่มเอกสารมาร่วมพิจารณา (Zeng, Wang and Nie, 2007)

ด้วยเหตุนี้งานวิจัยนี้จึงได้นำเสนอ การศึกษาเปรียบเทียบกระบวนการลดมิติเอกสารประกอบด้วยวิธีดัชนีความหมายภายใน กระบวนการพีแอลเอส และการเลือกคำสำคัญโดยวิธีรีเลแวนซี่สกอร์ ในการสร้างข้อมูลเข้าให้แก่โครงข่ายประสาทสำหรับการจำแนกเอกสาร โดยมีวัตถุประสงค์ คือ เปรียบเทียบผลกระทบของข้อมูลเข้าแต่ละแบบกับโครงข่ายประสาท ในด้านความถูกต้องในการจำแนกและเวลาที่ใช้ในการเรียนรู้ โดยข้อมูลที่ใช้คือเอกสารงานวิจัยทางชีวสารสนเทศ ในการสร้างแบบจำลองการจำแนกและการทดสอบประสิทธิภาพ

1.2 วัตถุประสงค์ของการวิจัย

1) เพื่อนำเสนอขั้นตอนวิธีการจำแนกเอกสารโดยใช้โครงข่ายประสาทโดยมีตัวแทนเอกสารที่สร้างจากวิธีการลดมิติเอกสาร ประกอบด้วยวิธีดัชนีความหมายภายใน กระบวนการพีแอลเอส และการเลือกคำสำคัญโดยวิธีรีเลแวนซี่สกอร์

2) เปรียบเทียบประสิทธิภาพของการลดมิติแต่ละแบบที่มีผลต่อการจำแนกเอกสาร โดยโครงข่ายประสาท

1.3 ประโยชน์ที่ได้รับจากการศึกษา เชิงทฤษฎี และเชิงประยุกต์

1) พัฒนาองค์ความรู้ของการจำแนกเอกสารโดยใช้หลักการของโครงข่ายประสาทเทียมและวิธีการลดมิติเอกสาร

2) สามารถเลือกวิธีการลดมิติที่เหมาะสมในการสร้างข้อมูลเข้าให้แก่โครงข่ายประสาทในการจำแนกเอกสารอย่างเหมาะสม

1.4 ขอบเขตการวิจัย

เพื่อให้บรรลุวัตถุประสงค์ของการทำวิจัย จึงมีการกำหนดขอบเขตงานดังนี้

1) จำแนกเอกสารโดยโครงข่ายประสาท โดยเอกสารที่ใช้ทดสอบคือเอกสารงานวิจัยทางชีวสารสนเทศ

2) การเปรียบเทียบประสิทธิภาพด้านความถูกต้องและเวลาการเรียนรู้ของโครงข่ายประสาทที่มีข้อมูลเข้าได้จากการลดมิติ ประกอบด้วยวิธีดัชนีความหมายภายใน กระบวนการพีแอลเอส และการเลือกคำสำคัญโดยวิธีรีเลแวนซี่สกอร์

1.5 วิธีการวิจัย

เพื่อให้การทำวิจัยสำเร็จตามแผนที่กำหนดจึงออกแบบวิธีการดำเนินการวิจัยดังนี้

1) ศึกษาเอกสารที่เกี่ยวข้องกับการจำแนกเอกสาร
2) ศึกษากระบวนการการแปลงเอกสารให้อยู่ในรูปแบบที่เหมาะสม (Preprocessing)
3) ศึกษารูปแบบของข้อมูลชีวสารสนเทศ
4) ทำการแปลงข้อมูลให้อยู่ในรูปพร้อมใช้งาน
5) สร้างแบบจำลองการจำแนกเอกสาร
6) ทำการทดลองและวิเคราะห์ผลที่ได้จากการทดลอง

1.6 อุปกรณ์ที่ใช้ในการวิจัย

1.6.1 ฮาร์ดแวร์ (Hardware)

- หน่วยประมวลผลกลางทำงานด้วยความเร็ว 2.27 GHz
- หน่วยความจำหลัก (RAM) ขนาด 2048 MB
- หน่วยความจำสำรอง (Hard Disk) ความจุ 250 GB

1.6.2 ซอฟต์แวร์ (Software)

- ระบบปฏิบัติการไมโครซอฟท์ วินโดวส์ วิสตา (Microsoft Windows Vista)
- โปรแกรมภาษาอาร์ (R) เวอร์ชั่น 2.6.1
- แพกเก็จทีเอ็ม (Text Mining: TM) ของภาษาอาร์
- โปรแกรมแมทแล็บ (Matlab) เวอร์ชั่น R2007a
- โปรแกรมไมโครซอฟท์ เวิร์ด 2007 (Microsoft Word 2007)
- โปรแกรมไมโครซอฟท์ เอ็กเซล 2007 (Microsoft Excel 2007)

1.7 สถานที่ที่ใช้ในการดำเนินการวิจัยและรวบรวมข้อมูล

1) ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่
2) สำนักหอสมุด มหาวิทยาลัยเชียงใหม่
3) ห้องปฏิบัติการวิจัยชีวสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่