การศึกษาทดลองเทคนิคการลดคุณลักษณะ และอัลกอริทึมการจัดหมวดหมู่ของเอกสารภาษาไทย

An Experimental Study on Feature Reduction Techniques and Classification Algorithms of Thai Documents

นิเวศ จิระวิชิตชัย* ปริญญา สงวนสัตย์ ** พยุง มีสัจ*

Nivet Chirawichitchai, Parinya Sanguansat and Phayung Meesad

*คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพมหานคร

**คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยรังสิต กรุงเทพมหานคร

บทคัดย่อ

งานวิจัยนี้ได้นำเสนอแบบจำลองการจัดหมวดหมู่เอกสารภาษาไทย โดยมุ่งเน้นวิธีการลดคุณลักษณะของเอกสารก่อนประมวลผลด้วยเครื่องจักรการเรียนรู้ เพื่อประโยชน์ในการลดระยะเวลาประมวลผลและประหยัดทรัพยากรของระบบ และเพิ่มประสิทธิภาพในการจัดหมวดหมู่เอกสารภาษาไทย จากการทดลองพบว่าการลดคุณลักษณะด้วยวิธี Chi-square ให้ประสิทธิภาพในการลดคุณลักษณะดีที่สุด รองลงมาเป็นการลดคุณลักษณะด้วย Information Gain และ Document Frequency ตามลำดับ เมื่อทดสอบประสิทธิภาพในการการจัดหมวดหมู่เอกสารพบว่าอัลกอริทึม Support Vector Machine มีประสิทธิภาพสูงที่สุด รองลงมาเป็นอัลกอริทึม Naïve-Bayes และอัลกอริทึม Decision Tree ตามลำดับ ผลทดลองการลดขนาดคุณลักษณะจากกลุ่มตัวอย่างพบว่าสามารถลดลงได้มากถึง 90% โดยการลดลงของคุณลักษณะดังกล่าวไม่ส่งผลให้ประสิทธิภาพในการจัดหมวดหมู่เอกสารลดลงแต่อย่างใด

คำสำคัญ : การลดคุณลักษณะ , การจัดหมวดหมู่เอกสาร , เครื่องจักรเรียนรู้

Abstract

This research presented thai document categorization framework focused on feature reduction before processing with machine learning to reduce the time and resources in the system of processing and increase the efficiency of thai document categorization. The results showed that Chi-square method was the most effective way to reduce the feature followed by Information Gain and Document Frequency, respectively. When testing the performance of categorization documents, it was found that the Support Vector Machine algorithms was the most effetive method

followed by Naïve-Bayes and Decision Tree algorithms, respectively. The sample size of feature was reduced up to 90% and the feature reduction does not affect the performance of document categorization.

Keywords : Feature Reduction, Document Categorization, Machine Learning

1. ความเป็นมาและความสำคัญของปัญหา

การขยายตัวทางการใช้งานระบบคอมพิวเตอร์และอินเตอร์เน็ต ตลอดช่วงระยะเวลาที่ผ่านมาจนถึงปัจจุบันมีแนวโน้มในการใช้งานเพิ่มมากขึ้นอย่างรวดเร็ว ส่งผลให้เกิดการสร้างและเก็บข้อมูลหลายชนิดในรูปแบบอิเล็กทรอนิกส์ ซึ่งหนึ่งในข้อมูลอิเล็กทรอนิกส์เหล่านี้คือข้อมูลประเภทเอกสาร เช่น จดหมายอิเล็กทรอนิกส์ (E-mail), เว็บเพจ (Web page) เอกสารข่าว (News) ไฟล์งานเอกสารต่าง (Document) ซึ่งเป็นข้อมูลที่มีปริมาณและเนื้อหาที่หลากหลายมากขึ้น ทำให้ยากต่อการค้นหาและจัดเก็บหมวดหมู่เอกสาร ดังนั้นการสืบค้นและการจัดการเอกสารจะง่ายและเป็นไปตามความต้องการต้องอาศัยการจัดแบ่งเอกสารเป็นกลุ่มหรือหมวดหมู่ให้สอดคล้องและตรงกับดัชนี เพื่อให้จัดเก็บและค้นคืนเอกสารได้อย่างรวดเร็วและมีประสิทธิภาพ จึงมีความจำเป็นต้องอาศัยผู้เชี่ยวชาญในการจัดกลุ่มข้อมูล ฉะนั้นจึงเป็นการยากในการที่จะจัดกลุ่มหรือแยกประเภทเอกสาร ยิ่งหากเอกสารมีปริมาณมากขึ้นทุกๆวัน ซึ่งต้องพึ่งพาทรัพยากรบุคคลในการจัดหมวดหมู่เอกสารเหล่านี้มากตามไปด้วยเช่นกัน ทำให้มีการคิดค้นพัฒนากระบวนการในการจัดหมวดข้อมูลที่มีขนาดใหญ่เหล่านี้ให้เป็นไปแบบอัตโนมัติ เพื่อที่จะสามารถจำแนกกลุ่มข้อมูลเพื่อใช้ประโยชน์จากข้อมูลและการจัดการกับข้อมูลให้มีประสิทธิภาพ รองรับการสืบค้นจากผู้ใช้งานเอกสารอย่างถูกต้องและเหมาะสม [1-2]

ปัจจุบันได้มีศึกษาเกี่ยวกับการนำวิธีการเรียนรู้ด้วยคอมพิวเตอร์ มาประยุกต์ร่วมกับการประมวลผลภาษาธรรมชาติเพื่อใช้จัดแบ่งกลุ่มเอกสารนั้น สามารถแบ่งได้ 2 ลักษณะ คือ การจัดกลุ่ม (Clustering) และการจำแนกหมวดหมู่ (Classification หรือ Categorization) การจัดกลุ่มเอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสารโดยไม่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน ซึ่งจะเป็นการแบ่งกลุ่มตามลักษณะของเอกสาร โดยเอกสารที่มีลักษณะเหมือนกันจะอยู่ด้วยกัน ส่วนการจำแนกหมวดหมู่เอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสาร โดยที่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน โดยจะเปรียบเทียบเอกสารกับต้นแบบในแต่ละหมวดหมู่เอกสารจะถูกจัดอยู่ในหมวดหมู่ที่ต้นแบบมีลักษณะคล้ายกับตัวมันเองมากที่สุด โดยผลลัพธ์ที่ได้จากวิธีการเรียนรู้ด้วยคอมพิวเตอร์นั้น ความถูกต้องใกล้เคียงกับผลการจำแนกหมวดหมู่ของเอกสารที่ทำโดยมนุษย์ ทำให้ประหยัดแรงงานมนุษย์เป็นอย่างมากเพราะไม่ต้องอาศัยผู้เชี่ยวชาญในการจำแนกประเภทเอกสารหรือปรับเปลี่ยนหมวดหมู่ของเอกสาร [1-2]

แต่เนื่องคุณลักษณะที่สกัดได้จากเอกสารที่จะใช้ในการเรียนรู้ เพื่อสร้างแบบจำลองในการจัดหมวดหมู่เอกสารนั้นมีปริมาณมาก ซึ่งคุณลักษณะจำนวนมากมายดังกล่าวนั้นจะต้องใช้ทรัพยากร

ของระบบและระยะเวลาในการประมวลผลจำนวนมาก รวมถึงอัลกอริทึมที่ใช้ในการเรียนรู้เพื่อสร้างตัวจำแนกหมวดหมู่โดยทั่วไปไม่สามารถรองรับการทำงานกับจำนวนคุณลักษณะของเอกสารที่สูงมากได้ดี จากความสำคัญของปัญหาดังกล่าวผู้วิจัยจึงมีแนวคิดที่จะแก้ปัญหาดังกล่าว โดยเสนอวิธีการการลดขนาดคุณลักษณะด้วยความถี่เอกสาร (Document Frequency) ค่าการเพิ่มของข้อมูล (Information Gain) และค่าสถิติไคแสควร์ (Chisquare) ก่อนส่งเข้าประมวลผลในส่วนของการเรียนรู้เอกสาร เพื่อลดระยะเวลาในการประมวลผลและทรัพยากรของระบบ [3-4] โดยทำการทดสอบประสิทธิภาพในการจัดหมวดหมู่เอกสารภาษาไทยกับอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ซัพพอร์ทเวกเตอร์แมชชีน(Support Vector Machine) เนอีฟเบย์(Naïve-Bayes) [5]

2. ทฤษฎีที่เกี่ยวข้อง

2.1 การตัดคำ (Word Segmentation)

การตัดคำ (Word Segmentation) การประมวลผลจำแนกหมวดหมู่เอกสารภาษาไทยได้อย่างมีประสิทธิภาพนั้น มีปัญหาเบื้องต้นคือ การตัดคำในภาษาไทย ซึ่งลักษณะการเขียนภาษาไทยจะมีการเขียนติดต่อกันเป็นสายอักขระโดยไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำ ดังเช่นภาษาอังกฤษซึ่งใช้ช่องว่าง(Space) คั่นระหว่างคำ ซึ่งเป็นอุปสรรคอย่างหนึ่งที่เพื่อแบ่งสายอักขระไทยออกเป็นคำๆ จึงได้มีผู้คิดค้นพัฒนาวิธีการตัดคำแบ่งได้เป็น หลักการตัดคำโดยใช้กฎ (Rule Base Approach) หลักการตัดคำโดยใช้อัลกอริทึม (Algorithm Approach) หลักการตัดคำโดยใช้พจนานุกรม (Dictionary Approach) และหลักการตัดคำโดยใช้คลังข้อมูล (CorpusBase Approach) แต่ละวิธีการต่างๆ ก็ให้ผลในด้านความถูกต้อง ความรวดเร็วของการทำงานและปริมาณการใช้ทรัพยากรต่างๆ ที่แตกต่างกัน จากการศึกษาเรื่องตัดคำสำหรับการจัดหมวดหมู่เอกสารภาษาไทย พบปัญหาด้านการหาขอบเขตของคำ เนื่องจากไม่มีการเขียนแบ่งพยางค์ คำ หรือประโยค ไม่มีหลักเกณฑ์ตายตัวในการใช้ช่องว่างในภาษาเขียน การสะกดคำมีรูปแบบซับซ้อน มีคำยืม คำทับศัพท์ คำเฉพาะจำนวนมาก และคำมีความกำกวมสูง จากการศึกษาเปรียบเทียบประสิทธิภาพวิธีดังกล่าวพบว่าวิธีตัดคำที่เหมาะสมกับการจัดหมวดหมู่เอกสารคือวิธีการตัดคำแบบยาวที่สุด (Longest Matching) ซึ่งมีวิธีการตรวจสอบสายอักขระ (String) ที่เข้ามาจากซ้ายไปขวากับพยางค์ที่เก็บไว้ในพจนานุกรม ในกรณีที่ตรวจสอบแล้วปรากฏว่าพบพยางค์มากกว่า 1 พยางค์ในพจนานุกรม ก็ให้เลือกแบ่งพยางค์โดยเลือกพยางค์ที่ยาวที่สุด แล้วทำต่อไปเรื่อยๆ จนจบสายอักขระ แต่ถ้ากรณีที่เลือกพยางค์ที่ยาวที่สุดแล้วทำให้เกิดพยางค์ที่ไม่ปรากฏในพจนานุกรมก็ยอมให้มีการย้อนรอย (Back Tracking) กลับไปเลือกพยางค์ที่ยาวรองมาแทนทำต่อไปเรื่อยๆจนสิ้นสุดสายอักขระ [6]

2.2 การกำจัดคำหยุด (Stop-Word List Removal)

เป็นการนำคำที่ไม่มีนัยสำคัญออกโดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลงคำที่ไม่มีนัยสำคัญ ในที่นี้หมายถึงคำที่ใช้กันโดยทั่วไปไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจาก

เอกสารแล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง ตัวอย่างเช่น คำบุพบทเป็นคำที่ใช้เชื่อมคำหรือกลุ่มคำให้สัมพันธ์กัน คำสันธานเป็นคำที่ทำหน้าที่เชื่อมคำกับคำ คำสรรพนามเป็นคำที่ใช้แทนคำนามที่กล่าวถึงมาแล้วในประโยค เป็นต้น คำหยุดมักเป็นคำที่ปรากฏขึ้นบ่อยครั้งในเอกสาร และปรากฏในเอกสารเกือบทุกฉบับ จึงถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการค้นคืนหรือการจำแนกหมวดหมู่ ดังนั้นการกำจัดคำหยุดจึงเป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี เพื่อกำจัดคุณลักษณะที่ไม่เป็นประโยชน์และลดขนาดของดัชนีลง ซึ่งจะช่วยประหยัดทั้งพื้นที่และเวลาในการประมวลผล [7-9]

2.3 การหารากศัพท์ (Stemming)

จึงเป็นการหารูปเดิมของคำ หรือหาคำที่มีความหมายคล้ายกัน เพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์เป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี ทำให้สามารถลดขนาดของดัชนีลงและเพิ่มประสิทธิภาพในการค้นคืนหรือการจำแนกหมวดหมู่ การหารากศัพท์ของคำภาษาไทยนั้นจะใช้วิธีการรวบรวมคำศัพท์ที่มีความหมายคล้ายกัน หรือมีรากศัพท์เดียวกันไว้เป็นรายการคำศัพท์ หรือจัดเก็บในคลังคำ เพื่อใช้ในการเปรียบเทียบหารากศัพท์ ซึ่งวิธีการนี้ต้องอาศัยมนุษย์เป็นผู้กำหนดไว้ก่อนว่า คำแต่ละคำมีรากศัพท์ เป็นคำใด วิธีการนี้ต้องอาศัยผู้เชี่ยวชาญทางภาษาและใช้เวลาในการเก็บรวบรวมและจัดทำรายการคำศัพท์ [7-8]

2.4 การสกัดคุณลักษณะ (Feature Extraction)

วัตถุประสงค์ของขั้นตอนการสกัดคุณลักษณะเอกสารคือการดึงคุณลักษณะ (Feature) ของเอกสารออกมา กับการลดขนาดเอกสารลง ซึ่งการดึงคุณลักษณะออกมานั้น ก่อนอื่นเราต้องการกำหนดก่อนว่าจะใช้อะไร เป็นตัวแทนคุณลักษณะของเอกสาร และใช้ค่าใดแทนคุณลักษณะเอกสารนั้น จากการสำรวจงานวิจัยที่ผ่านมาทั้งในประเทศและต่างประเทศพบว่า ส่วนใหญ่จะใช้คำเป็นตัวแทนคุณลักษณะของเอกสาร และใช้พื้นฐานค่าความถี่ของคำเป็นค่าของคุณลักษณะ นอกจากการใช้คำเดี่ยวแล้ว ยังสามารถใช้ วลี หรือกลุ่มของคำ ประโยค ฯ ล ฯ แทนคุณลักษณะของเอกสารได้เช่นกัน ตัวแทนคุณลักษณะของเอกสารที่นิยมใช้ในการจัดหมวดหมู่เอกสารประเภทข้อความคือ ถุงคำ (Bag of words) ซึ่งจะเก็บอยู่ในรูปแบบของเวกเตอร์ โดยองค์ประกอบของเวกเตอร์อาจจะแทนด้วยคุณลักษณะของค่าความจริง (Boolean features) แทนด้วยค่าความถี่ หรือแทนด้วยค่าน้ำหนักของคำ ค่าน้ำหนักของชุดลำดับคำ (series of words, n-gram) หรือค่าน้ำหนักของวลีในเอกสาร [5] โดยการเลือกหน่วยคุณลักษณะที่จะมาใช้เป็นตัวแทนเอกสาร ซึ่งงานวิจัยนี้ใช้คุณลักษณะแบบคำเดียว (Sigle word)

2.5 การสร้างดัชนี (indexing)

เนื่องจากคอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ของเอกสารซึ่งเป็นภาษาธรรมชาติโดยตรงได้ดังนั้นจึงต้องแปลงเอกสารให้อยู่ในรูปแบบ ที่คอมพิวเตอร์สามารถใช้ในการเรียนรู้ได้ ขั้นตอนในการแปลงเอกสาร เรียกว่า การทำดัชนี (Indexing) เพื่อสร้างตัวแทนเนื้อหาของเอกสาร (Document

Representation) สำหรับใช้ในกระบวนการเรียนรู้ วัตถุประสงค์ของการสร้างดัชนีคือ การคำนวณหาค่าที่จะมาใช้เป็นค่าคุณลักษณะของเอกสาร หรืออาจจะเรียกได้ว่าการหาค่าน้ำหนัก (term weighting) การสร้างดัชนี โดยทั่วไปที่นิยมใช้กัน จะเริ่มจากการสร้างเวกเตอร์ตัวแทนเอกสาร จากนั้นจะสร้างเมตริกซ์ของกลุ่มเอกสารขึ้นจากเวกเตอร์เอกสารทั้งหมดในกลุ่ม [1,5] ซึ่งงานวิจัยนี้ใช้วิธีความถี่ของคำที่ปรากฏในเอกสารมาเป็นค่าน้ำหนัก ถ้าคำใดมีความถี่มากก็จะได้ค่านำหนักที่มีค่าสูงมากตาม

2.6 การเลือกคุณลักษณะ (Feature Selection)

จากที่กล่าวมาจะพบว่าเอกสารนั้นมีแนวโน้มที่จะเพิ่มปริมาณสูงขึ้นทุกวัน ทำให้เอกสารมีจำนวนคุณลักษณะมากขึ้น ทำให้วิธีเบื้องต้นในการลดขนาดเอกสารคือการใช้การนำคำที่ไม่มีนัยสำคัญออกกับการทำรากศัพท์แล้วยังไม่เพียงพอ ซึ่งจำนวนคุณลักษณะมีผลต่อประสิทธิภาพของการจำแนกหมวดหมู่เอกสาร เนื่องจากอัลกอริทึมที่ใช้ในการเรียนรู้เพื่อสร้างตัวจำแนกหมวดหมู่โดยทั่วไปไม่สามารถรองรับการทำงานกับจำนวนคุณลักษณะของเอกสารที่สูงมากได้ดี การลดขนาดเอกสารจึงเป็นขั้นตอนหนึ่งที่จะต้องทำก่อน การสร้างตัวจำแนกเอกสารแต่การลดขนาดของเอกสารต้องพิจารณาด้วยความระมัดระวัง เนื่องจากมีความเสี่ยงในการที่จะกำจัดคุณลักษณะที่สำคัญต่อการจำแนกหมวดหมู่ออกไป จากการศึกษาพบว่าวิธีการลดคุณลักษณะเอกสารประกอบด้วยการสร้างคุณลักษณะใหม่จากคุณลักษณะเดิม อาจจะนำคุณลักษณะพื้นฐานเหล่านี้มารวมกันเพื่อให้เป็นคุณลักษณะใหม่ที่มีระดับสูงขึ้นค่าที่นิยมใช้ได้แก่ค่า [3-5,8,10]

ความถี่เอกสาร (DF: Document Frequency) ค่า DF คำนวณได้จากการนับจำนวนเอกสารที่ปรากฏคำที่ต้องการอยู่ การลดขนาดเอกสารทำได้โดยจากดูจากค่า DF เช่นกำหนดว่า ถ้าคำใดมีค่าDF ต่ำกว่าเกณฑ์ ให้ตัดออก เป็นต้น

ค่าการเพิ่มของข้อมูล (IG: Information Gain) ค่า IG จะถูกใช้บ่อยในงานที่เกี่ยวข้องกับเครื่องจักรการเรียนรู้ จะใช้เป็นเกณฑ์ในการหาค่าความดีของเทอม ค่าจะคำนวณจากจำนวนบิตที่ได้รับสำหรับการทำนายกลุ่มโดยการดูจากการมีอยู่หรือไม่มีอยู่ของคำในเอกสาร ให้ $C_1, ..., C_K$ แทนเซ็ตที่เป็นไปได้ของกลุ่ม ค่า IG ของคำ w นิยามโดย [8]

$$IG(w) = -\sum_{j=1}^{K} P(c_j)\log P(c_j) + P(w)\sum_{j=1}^{K} P(c_j \mid w)\log P(c_j \mid w) - P(\overline{w})\sum_{j=1}^{K} P(c_j \mid \overline{w})\log P(c_j \mid \overline{w})$$

ค่า $P(C_j)$ คำนวณได้จาก เศษส่วนของจำนวนเอกสารที่อยู่กลุ่ม C_j กับ จำนวนเอกสารทั้งหมด

ค่า P(w) คำนวณได้จาก เศษส่วนของจำนวนเอกสารที่มีคำ w กับจำนวนเอกสารทั้งหมด

ค่า $P(C_j \mid w)$ คำนวณได้จาก เศษส่วนของคำจำนวนเอกกลุ่ม C_j ที่มีคำ w กับเอกสารทั้งหมด

ค่า $P(C_j \mid \overline{w})$ คำนวณได้จาก เศษส่วนของคำจำนวนเอกกลุ่ม C_j ที่ไม่มีคำ w กับเอกสารทั้งหมด

ค่าสถิติไคสแควร์ (Chisquare) ค่าสถิติไคสแควร์ วัดความเป็นอิสระต่อกันระหว่างคำ และกลุ่มนิยามโดย [8]

$$\chi^2(w,c_j) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

A คือ จำนวนเอกสารจากกลุ่ม C_j ซึ่งมีคำ w

B คือ จำนวนเอกสารซึ่งมีคำ w แต่ไม่ได้อยู่ในกลุ่ม C_j

C คือ จำนวนเอกสารจากกลุ่ม C_j ซึ่งไม่มีคำ w

D คือ จำนวนเอกสารที่ไม่ได้อยู่ในกลุ่ม C_j หรือไม่มีคำ w

N คือ จำนวนเอกสารทั้งหมด

ค่าสถิติไคสแคว์ที่ใช้ในการเลือกคือ ค่าเฉลี่ย กับค่าสูงสุด

$$\chi^2_{avg}(w) = \sum_{j=1}^{L} P(c_j)\chi^2(w,c_j)$$

$$\chi^2_{max}(w) = \max_j \chi^2(w,c_j)$$

2.7 อัลกอริทึมการจัดหมวดหมู่ (Classifier Algorithm)

อัลกอริทึมในการจัดหมวดหมู่การเรียนรู้แบบมีผลเฉลย (Supervised Learning) สามารถแบ่งขั้นตอนวิธีการจัดหมวดหมู่เอกสารแบ่งได้เป็น 2 ขั้นตอนคือ การเรียนรู้เพื่อสร้างกลุ่มเอกสารต้นแบบและแยกหมวดหมู่ของเอกสารที่สนใจ โดยการตรวจสอบหาความคล้ายกับกลุ่มเอกสารต้นแบบ [1]

2.7.1 ต้นไม้ตัดสินใจ (Decision Tree) ต้นไม้จะประกอบด้วยโหนดแทนคุณลักษณะ และโหนดล่างสุดแทนหมวดหมู่ การสร้างกิ่งสาขาจะพิจารณาจากค่าความจริงของคุณลักษณะ โดยค่าความจริงที่ใช้จะมาจากการคำนวณหาค่า Entropy และค่า Information Gain คุณลักษณะใดที่มีค่า Information Gain มากที่สุดจะถูกเลือกเป็นโหนดลูก หากมีค่า Entropy เป็น 0 จะได้เป็นโหนดล่างสุด

2.7.2 ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) หลักการของ SVM คือการสร้างสมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกันโดย SVM จะพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มี ระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มมากที่สุด SVM จะใช้ฟังก์ชั่นแม็ปสำหรับย้ายข้อมูลจาก Input Space ไปยัง Feature Space และสร้างฟังก์ชั่นวัดความคล้ายที่เรียกว่า (Kernel Function) บน Feature Space

2.7.3 เนอีฟเบย์ (Naïve-Bayes) ใช้ทฤษฎี Bayes Theorem ในการคำนวณความน่าจะเป็นซึ่งถูกใช้ในการทำนายผล Naïve-Bayes เป็นเทคนิคในการแก้ปัญหาแบบ classification ที่ทั้งสามารถ

คาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ด้วย มันจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์

3. วิธีการดำเนินการวิจัย

งานวิจัยนี้ทำการทดลองการลดคุณลักษณะร่วมกับอัลกอริทึมการจัดหมวดหมู่เอกสาร โดยมุ่งเน้นจัดหมวดหมู่เอกสารภาษาไทย โดยทดสอบกับเอกสารประเภทข่าวอิเล็กทรอนิกส์จากหนังสือพิมพ์ไทยรัฐ จำนวน 10 กลุ่ม ได้แก่ กลุ่มการศึกษา กลุ่มบันเทิง กลุ่มสังคม กลุ่มการเมือง กลุ่มเทคโนโลยี กลุ่มกีฬา กลุ่มข่าวต่างประเทศ กลุ่มเกษตร กลุ่มเศรษฐกิจ กลุ่มวัฒนธรรม โดยมีจำนวนกลุ่มตัวอย่างทั้งหมด 2000 เอกสาร เป็นกลุ่มตัวอย่างเรียนรู้ และ ทำการทดสอบด้วยวิธี 10-fold cross validation โดยแบบจำลองที่นำเสนอในงานวิจัยนี้ จะทำการตัดคำโดยวิธีการตัดคำแบบยาวที่สุด (Longest Matching) โดยใช้พจนานุกรมฉบับ Lexitron และทำการกำจัดคำหยุดและทำรากศัพท์จากฐานข้อมูลที่กำหนดขึ้น หลังจากนั้นทำการลดขนาดคุณลักษณะด้วย วิธีความถี่เอกสาร (DF) ค่าการเพิ่มของข้อมูล (IG) ค่าสถิติไคแสควร์ (Chisquare) ซึ่งวิธีการลดขนาดคุณลักษณะดังกล่าวเป็นวิธีที่ง่ายและใช้เวลาการประมวลผลน้อยแต่สามารถคัดเลือกคุณลักษณะที่มีนัยสำคัญได้ดี ผลที่ได้จากขั้นตอนดังกล่าวจะทำการลดขนาดของคุณลักษณะของเอกสารลง แล้วจะถูกส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลยด้วยอัลกอริทึม ต้นไม้ตัดสินใจ (Decision Tree) เนอีฟเบย์ (Naïve-Bayes) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) มาเรียนรู้ ทำการทดสอบเปรียบเทียบประสิทธิภาพด้านความถูกต้อง ความแม่นยำ โดยใช้วิธีการประเมินความสามารถของแบบจำลอง โดยวัดที่ประสิทธิผลของการจำแนกหมวดหมู่ตามแนวคิดทางด้านการค้นคืนสารสนเทศ ซึ่งก็คือการวัดค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) และค่า F-Measurement ซึ่งคำนวณได้ดังสมการ [8]

$$recall = \frac{a}{a+c}$$

$$precision = \frac{a}{a+b}$$

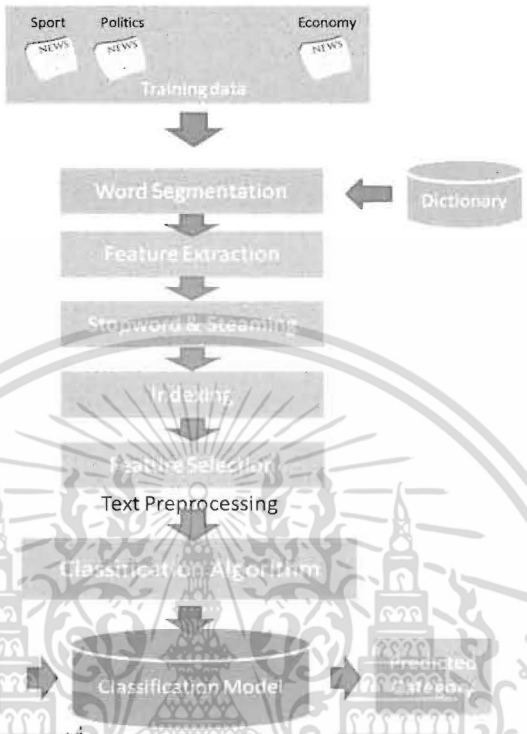
$$F\text{-}measure = \frac{2 \times (Precision \times Recall)}{Precision + Recall}$$

โดยให้ a = เป็นจำนวนเอกสารที่จัดกลุ่มถูก (correctly assigned)

b = เป็นจำนวนเอกสารที่จัดกลุ่มถูก แต่ความจริงผิดกลุ่ม (incorrectly assigned)

c = เป็นจำนวนเอกสารที่จัดกลุ่มผิด แต่ความจริงถูกกลุ่ม (incorrectly rejected)

d = เป็นจำนวนเอกสารที่จัดกลุ่มผิด (correctly rejected)



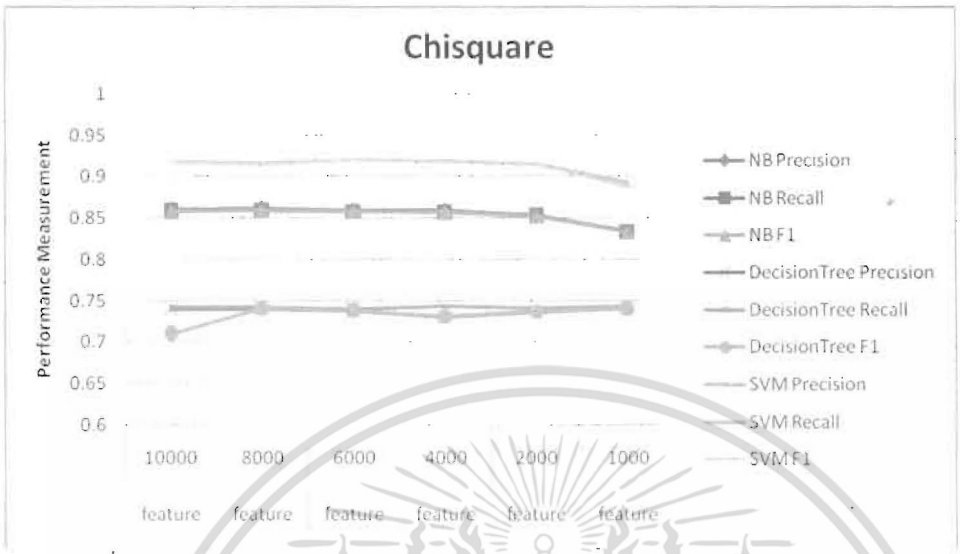
รูปที่ 1 แบบจำลองการจัดหมวดหมู่เอกสาร

4. ผลการทดลอง

การทดลองการลดคุณลักษณะร่วมกับอัลกอริทึมการจัดหมวดหมู่เอกสารซึ่งประกอบด้วย อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) เนอีฟเบย์ (Naïve-Bayes) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) โดยทดลองกับกลุ่มตัวอย่างเอกสารข่าวภาษาไทยจำนวน 10 ประเภท จำนวน 2000 เอกสาร ซึ่งมีการกระจายตัวของกลุ่มตัวอย่างเท่ากันคือประเภทละ 200 เอกสาร ซึ่งได้ผลการทดลองดังนี้

ตารางที่ 1 ผลการทดลองลดคุณลักษณะด้วยวิธี chi-square

ChiSquare	NB			DecisionTree			SVM		
feature	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
10000	0.861	0.859	0.858	0.74	0.742	0.710	0.918	0.916	0.916
8000	0.862	0.859	0.859	0.74	0.742	0.741	0.917	0.915	0.915
6000	0.86	0.858	0.858	0.738	0.74	0.738	0.921	0.92	0.920
4000	0.86	0.858	0.857	0.73	0.743	0.731	0.919	0.918	0.918
2000	0.854	0.853	0.852	0.736	0.741	0.737	0.916	0.915	0.915
1000	0.835	0.834	0.833	0.741	0.744	0.741	0.893	0.888	0.888
average	0.855	0.854	0.853	0.7375	0.742	0.733	0.914	0.912	0.912

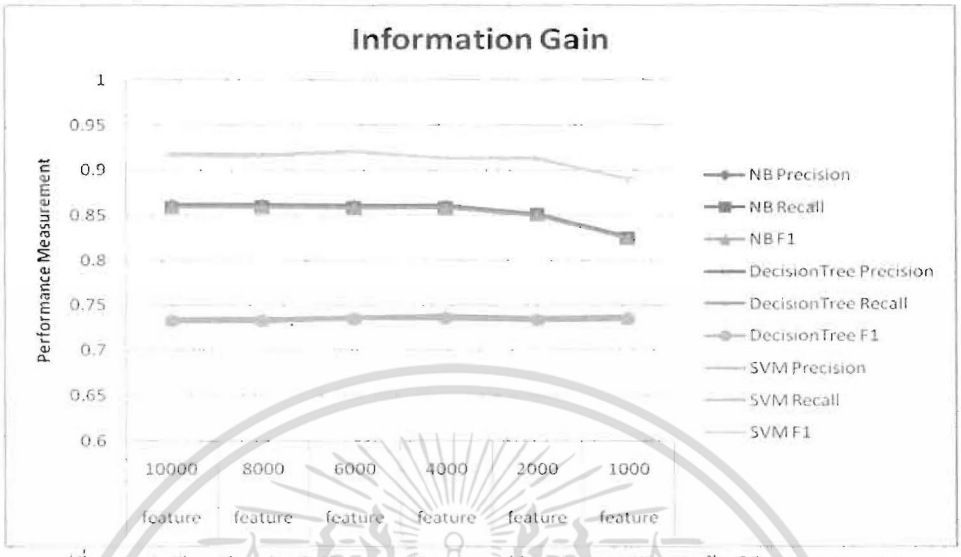


รูปที่ 2 กราฟเปรียบเทียบประสิทธิภาพการจัดหมวดหมู่ร่วมกับลดคุณลักษณะด้วยวิธี chi-square

จากการทดลองโดยใช้ chi-square ในการลดคุณลักษณะและทำการเรียนรู้ โดยวัดจากค่า F-Measurement สามารถสรุปได้ว่า อัลกอริทึม Support Vector Machine ให้ประสิทธิภาพการจัดหมวดหมู่โดยเฉลี่ยออกมาดีที่สุดคือ 91.2 % รองลงมาเป็นอัลกอริทึม Naïve-Bayes ให้ประสิทธิภาพการจัดหมวดหมู่โดยเฉลี่ย 85.3 % และตัวสุดท้ายคืออัลกอริทึม Decision Tree 73.3 % เมื่อตรวจสอบการลดคุณลักษณะที่ได้จากการทดลองโดยเทียบจาก 10000 คุณลักษณะเป็นตัวแบบ พบว่าอัลกอริทึม Support Vector Machine สามารถลดคุณลักษณะเหลือ 2000 โดยค่า F1 ยังคงซึ่งประสิทธิภาพเท่าตัวแบบ คือ 91% สามารถลดคุณลักษณะลงถึง 80 % ส่วนอัลกอริทึม Naïve-Bayes สามารถลดคุณลักษณะเหลือ 2000 เช่นกัน โดยมีประสิทธิภาพที่ 85% ส่วนอัลกอริทึม Decision Tree สามารถลดคุณลักษณะได้มากที่สุดเหลือ 1000 ลดคุณลักษณะลงถึง 90 % ในขณะที่ให้ประสิทธิภาพการจัดหมวดหมู่สูงกว่าตัวแบบ โดยมีประสิทธิภาพที่ 74% มากกว่าตัวแบบ 3 %

ตารางที่ 2 ผลการทดลองลดคุณลักษณะด้วยวิธี Information Gain

IG	NB			DecisionTree			SVM		
feature	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
10000	0.862	0.859	0.859	0.732	0.735	0.733	0.918	0.916	0.916
8000	0.862	0.860	0.859	0.732	0.735	0.733	0.917	0.915	0.915
6000	0.861	0.859	0.858	0.735	0.737	0.735	0.921	0.920	0.920
4000	0.861	0.859	0.858	0.735	0.739	0.736	0.913	0.913	0.913
2000	0.852	0.850	0.850	0.733	0.736	0.734	0.913	0.912	0.912
1000	0.826	0.825	0.824	0.734	0.738	0.735	0.890	0.890	0.890
average	0.854	0.852	0.851	0.734	0.737	0.734	0.912	0.911	0.911

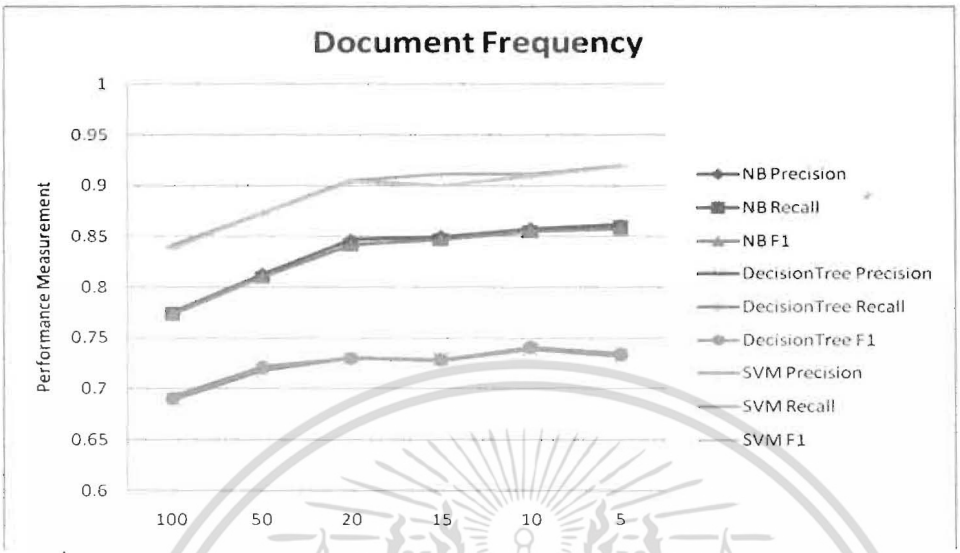


รูปที่ 3 กราฟเปรียบเทียบประสิทธิภาพการจัดหมวดหมู่ร่วมกับลดคุณลักษณะด้วยวิธี Information Gain

ส่วนการทดลองโดยใช้ Information Gain ในการลดคุณลักษณะและทำการเรียนรู้ โดยวัดจากค่า F-Measurement สามารถสรุปได้ว่า อัลกอริทึม Support Vector Machine ให้ประสิทธิภาพการจัดหมวดหมู่โดยเฉลี่ยออกมาดีที่สุดคือ 91.1 % รองลงมาเป็นอัลกอริทึม Naïve-Bayes ให้ประสิทธิภาพการจัดหมวดหมู่โดยเฉลี่ย 85.1 % และตัวสุดท้ายคืออัลกอริทึม Decision Tree 73.4 % เมื่อตรวจสอบการลดคุณลักษณะที่ได้จากการทดลองโดยเทียบจาก 10000 คุณลักษณะเป็นตัวแบบ พบว่าอัลกอริทึม Support Vector Machine สามารถลดคุณลักษณะเหลือ 2000 โดยค่า F1 ยังคงซึ่งประสิทธิภาพเท่าตัวแบบ คือ 91% สามารถลดคุณลักษณะลงถึง 80 % ส่วนอัลกอริทึม Naïve-Bayes สามารถลดคุณลักษณะเหลือ 2000 เช่นกัน โดยมีประสิทธิภาพที่ 85% ส่วนอัลกอริทึม Decision Tree สามารถลดคุณลักษณะได้มากที่สุดเหลือ 1000 ในขณะที่ยังคงให้ประสิทธิภาพการจัดหมวดหมู่เท่ากับตัวแบบโดยมีประสิทธิภาพที่ 73%

ตารางที่ 3 ผลการทดลองลดคุณลักษณะด้วยวิธี Document Frequency

	NB			DecisionTree			SVM		
DF	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
100	0.775	0.773	0.773	0.689	0.692	0.690	0.841	0.839	0.838
50	0.813	0.810	0.810	0.718	0.722	0.720	0.873	0.873	0.872
20	0.847	0.842	0.842	0.730	0.730	0.730	0.905	0.905	0.904
15	0.850	0.847	0.847	0.728	0.729	0.728	0.911	0.900	0.900
10	0.858	0.855	0.855	0.739	0.741	0.740	0.911	0.910	0.909
5	0.862	0.859	0.858	0.732	0.735	0.733	0.920	0.919	0.919
average	0.834	0.831	0.831	0.723	0.725	0.724	0.894	0.891	0.890

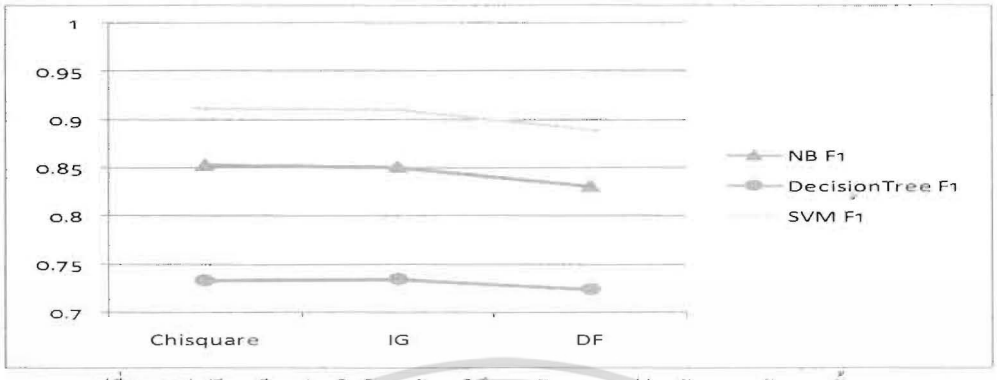


รูปที่ 4 กราฟเปรียบเทียบประสิทธิภาพการจัดหมวดหมู่ร่วมกับลดคุณลักษณะด้วยวิธี Document Frequency

ส่วนการทดลองโดยใช้ Document Frequency ในการลดคุณลักษณะและทำการเรียนรู้ โดยวัดค่า F-Measurement สามารถสรุปได้ว่าอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพการจัดหมวดหมู่โดยเฉลี่ยออกมาดีที่สุดคือ 89.0 % รองลงมาเป็นอัลกอริทึม Naïve-Bayes ให้ประสิทธิภาพการจัดหมวดหมู่โดยเฉลี่ย 83.1 % และตัวสุดท้ายคืออัลกอริทึม Decision Tree 72.4 % เมื่อตรวจสอบการลดคุณลักษณะโดยวัดจากค่า DF (ค่าความถี่เอกสารที่ปรากฏคำนั้น) พบว่าการลดคุณลักษณะแบบนี้ทำให้ทุกอัลกอริทึมมีประสิทธิภาพในการเรียนรู้จัดหมวดหมู่ลดลงตามค่า DF ที่มากขึ้น โดยพบว่าอัลกอริทึม Support Vector Machine ค่า DF เท่ากับ 5 ให้ประสิทธิภาพดีที่สุดคือ 91 % ส่วนอัลกอริทึม Naïve-Bayes ค่า DF เท่ากับ 5 ให้ประสิทธิภาพดีที่สุดคือ 85% และอัลกอริทึม Decision Tree ค่า DF เท่ากับ 10 ให้ประสิทธิภาพดีที่สุดคือ 74%

ตารางที่ 4 ผลการทดลองการลดคุณลักษณะร่วมกับอัลกอริทึมการจัดหมวดหมู่ทั้ง 3 แบบ

	NB			DecisionTree			SVM		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Chisquare	0.855	0.854	0.853	0.7375	0.742	0.733	0.914	0.912	***0.912***
IG	0.854	0.852	0.851	0.734	0.737	0.734	0.912	0.911	***0.911***
DF	0.834	0.831	0.831	0.723	0.725	0.724	0.894	0.891	***0.890***



รูปที่ 5 กราฟเปรียบเทียบประสิทธิภาพอัลกอริทึมการจัดหมวดหมู่ร่วมกับลดคุณลักษณะทั้ง 3 แบบ

เมื่อทำการทดลองโดยเปรียบเทียบการลดคุณลักษณะและทำการเรียนรู้ โดยวัดค่าประสิทธิภาพด้วย F-Measurement สามารถสรุปได้ว่าการลดคุณลักษณะด้วย chi-square และเรียนรู้ด้วยอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพการจัดหมวดหมู่โดยเฉลี่ยออกมาดีที่สุดคือ 91.2 % รองลงมาเป็นการลดคุณลักษณะด้วย Information Gain และเรียนรู้ด้วยอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพการจัดหมวดหมู่โดยเฉลี่ย 91.1 % และสุดท้ายคือการลดคุณลักษณะด้วย Document Frequency และเรียนรู้ด้วยอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพการจัดหมวดหมู่โดยเฉลี่ย 89.0 %

5. สรุปผลและข้อเสนอแนะ

งานวิจัยนี้ได้นำเสนอวิธีการลดคุณลักษณะร่วมกับอัลกอริทึมเครื่องจักรการเรียนรู้ เพื่อศึกษาวิธีการลดคุณลักษณะที่เหมาะสมและมีประสิทธิภาพในการจัดหมู่เอกสารข่าวภาษาไทย จากการทดลองพบว่าการลดคุณลักษณะด้วยวิธี Chi-square ให้ประสิทธิภาพในการลดคุณลักษณะดีที่สุด รองลงมาเป็นการลดคุณลักษณะด้วย Information Gain และ Document Frequency ตามลำดับ เหตุผลที่ใช้วิธีการลดขนาดคุณลักษณะทั้ง 3 วิธีดังกล่าว อันเนื่องมาจากเป็นวิธีที่ง่าย ไม่ซับซ้อนและใช้เวลาการประมวลผลน้อย แต่สามารถคัดเลือกคุณลักษณะที่มีนัยสำคัญได้ดี และ Chi-square ให้ประสิทธิภาพดีที่สุด เนื่องจากวัดค่าความเป็นอิสระต่อกันระหว่างคำจากกลุ่มตัวอย่างโดยใช้เทคนิค Filtering เลือกคุณลักษณะที่ให้ค่าความสำคัญทางสถิติมากที่สุดเรียงตามลำดับ ซึ่งคุณลักษณะที่มีความสำคัญลำดับแรกๆจะมีอำนาจในการจำแนกกลุ่มเอกสารมากกว่าอันดับท้ายๆ เมื่อพิจารณาประสิทธิภาพด้านอัลกอริทึมเครื่องจักรการเรียนรู้ พบว่าอัลกอริทึม Support Vector Machine มีประสิทธิภาพในการการจัดหมวดหมู่เอกสารดีที่สุด รองลงมาเป็นอัลกอริทึม Naïve-Bayes และอัลกอริทึม Decision Tree ตามลำดับ อันเนื่องมาจากอัลกอริทึม Support Vector Machine มีพฤติกรรมที่จะแยกแยะข้อมูลโดยใช้สมการระนาบหลายมิติ โดยจะพยายามหาจุดข้อมูลที่ทำให้ได้

สมการระนาบหลายมิติที่ใช้แบ่งแยกดีที่สุด (Optimal Hyperplane) ความถูกต้องที่สุด โดยพิจารณาจากระยะห่าง (Margin) ระหว่างคลาส ซึ่งเส้นระนาบที่ดีที่สุดนี้จะสามารถจำแนกกลุ่มเอกสารออกมาได้อย่างมีประสิทธิภาพ ผลจากการทดลองลดขนาดคุณลักษณะจากกลุ่มตัวอย่าง พบว่าสามารถลดคุณลักษณะลงได้มากถึง 90 % โดยการลดลงของคุณลักษณะดังกล่าวไม่ส่งผลให้ประสิทธิภาพในการจัดหมวดหมู่เอกสารลดลงแต่อย่างใด แต่สามารถลดทรัพยากรของระบบและลดระยะเวลาในการประมวลผลได้เป็นอย่างมาก จากผลการทดลองนี้สามารถนำแบบจำลองนี้ไปประยุกต์กับใช้ประโยชน์ในการสร้างระบบจัดหมวดหมู่เอกสารอัตโนมัติและสามารถนำมาประยุกต์ใช้กับงานด้านอื่นๆ เช่น การคัดกรองเอกสาร (Document Filtering) การจัดทำดัชนีอัตโนมัติเพื่อใช้ในการค้นคืนเอกสาร (Automatic Indexing for IR System) การจัดหมวดหมู่ของเว็บเพจ (Web Page Classification) เป็นต้น

กิตติกรรมประกาศ

ขอขอบพระคุณ ดร.ชูชาติ หฤไชยะศักดิ์ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ที่ให้คำแนะนำและอนุเคราะห์กลุ่มตัวอย่างที่ใช้ในการทดลองงานวิจัยนี้

เอกสารอ้างอิง

[1] Sebastiani, Fabrizio. "Machine Learning in Automated Text Categorization". **ACM Computing Surveys (CSUR)**. March. 2002

[2] Marquez, Llus. "Machine learning and natural language processing". **Technical Report Departament de Llenguatges Sistemes Informatics (LSI)**, Barcelona, Spain. 2000

[3] Yang, Perderson, O. "A Comparative Study on Feature Selection in Text Categorization". Proceedings of ICML-97, 14th International Conference on Machine Learning. 1997

[4] Haruechaiyasak, C., Jitkrittum, W. , Sangkeettrakarn, C., Damrongrat, C. "Implementing News Article Category Browsing Based on Text Categorization Technique", International Conference on Web Intelligence and Intelligent Agent Technology. 2008

[5] Aas., Eikvil. "Text Categorization: A Survey". **Report Norwegian Computing Center**. 1999

[6] Charoenpornsawat ,P. "Feature-based Thai Word Segmentation". Master's Thesis. Computer Engineering. Chulalongkorn University, Bangkok, Thailand. 1999

[7] Jaruskulchai, C. "An Automatic Indexing for Thai Text Retrieval". PhD Thesis, George Washington University. USA. 1998

[8] วัลลภ อินทร์ฉ่ำ. ระบบการจัดหมวดหมู่เอกสารภาษาไทยอัตโนมัติโดยใช้ SVM ร่วมกับการประมวลผลภาษา,วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต มหาวิทยาลัยเกษตรศาสตร์. 2548

[9] ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC). "ศัพท์ที่พบในคลังข้อมูล". http://203.185.132.59/thailang/thaichar/word_thai.php

[10] Taira Hirotoshi, Haruno Masahiko. "Feature Selection in SVM Text Categorization", **Transactions of Information Processing Society of Japan**. 2000