แบบจำลองการจำแนกเอกสารภาษาไทยอัตโนมัติ

นิเวศ จิระวิชิตชัย*

บทคัดย่อ

บทความนี้เป็นการนำเสนอวิธีการสร้างแบบจำลองการจำแนกเอกสารภาษาไทยอัตโนมัติ เพื่อประโยชน์ในการแก้ปัญหาการจำแนกเอกสารที่มีปริมาณมากและช่วยประหยัดแรงงานมนุษย์เพราะไม่ต้องใช้มนุษย์ในการจำแนกเอกสาร ซึ่งขั้นตอนการสร้างแบบจำลองเอกสารประกอบด้วย 1) การสกัดคุณลักษณะด้วยการตัดคำ 2) การกำจัดคำหยุดและทำรากศัพท์ 3) การกำหนดค่าน้ำหนัก ดัชนี คำและการลดคุณลักษณะ และ 4) การเรียนรู้ด้วยเครื่องจักรการเรียนรู้แบบมีผู้สอนและทำการทดสอบประสิทธิภาพด้านความถูกต้องของแบบจำลองในจำแนกเอกสาร

คำสำคัญ : การจำแนกเอกสาร, แบบจำลอง, เหมืองข้อมูล

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีปทุม วิทยาเขตชลบุรี
*ผู้ติดต่อ,อีเมล์ : nivet99@hotmail.com รับเมื่อ 29 มีนาคม 2555 ตอบรับเมื่อ 10 กรกฎาคม 2555

Automatic Thai Document Classification Model

Nivet Chirawichitchai*

Abstract

This article presents the application of modeling to automatic classification of Thai document. Modeling is beneficial for solving the problem of classifying electronic documents in a large volume and save human labor. The process of modeling consists of 1) feature extraction with the Thai word segmentation 2) stop-word list removal and stemming. 3) the index of the documents and feature reduction and 4) learning by machine learning and testing the accuracy of the document classification model.

Keywords : Document classification, Modeling, Data mining

Graduate School, Sripatum University Chonburi Campus.

*Corresponding author, E-mail: nivet99@hotmail.com Received 29 March 2012; Accepted 10 July 2012

1. บทนำ

การขยายตัวทางการใช้งานระบบคอมพิวเตอร์และอินเทอร์เน็ตตลอดช่วงระยะเวลาที่ผ่านมาจนถึงปัจจุบันมีแนวโน้มในการใช้งานเพิ่มมากขึ้นอย่างรวดเร็ว ส่งผลให้เกิดการสร้างและเก็บข้อมูลหลายชนิดในรูปแบบอิเล็กทรอนิกส์ ซึ่งหนึ่งในข้อมูลอิเล็กทรอนิกส์เหล่านี้คือข้อมูลประเภทเช่น จดหมายอิเล็กทรอนิกส์ (E-mail) เว็บเพจ (Web page) เอกสารข่าว (News) ไฟล์งานเอกสารต่างๆ (Document) ซึ่งเป็นข้อมูลที่มีปริมาณและเนื้อหาที่หลากหลายมากขึ้น ทำให้ยากต่อการค้นหาและจัดเก็บหมวดหมู่เอกสาร ซึ่งมีความจำเป็นต้องใช้ผู้เชี่ยวชาญในการจัดกลุ่ม ดังนั้นจึงเป็นการยากในการที่จะจัดกลุ่มหรือแยกประเภทเอกสารยิ่งถ้าหากเอกสารมีปริมาณมากขึ้นทุกวัน ความต้องการของทรัพยากรบุคคลในการจำแนกเอกสารเหล่านี้ต้องมีมากตามไปด้วยเช่นกัน ทำให้มีการคิดค้นพัฒนากระบวนการในการจำแนกข้อมูลที่มีขนาดใหญ่เหล่านี้ ให้เป็นไปแบบอัตโนมัติ เพื่อที่จะสามารถจำแนกกลุ่มข้อมูล เพื่อใช้ประโยชน์จากข้อมูลและการจัดการกับข้อมูลให้มีประสิทธิภาพ รองรับการสืบค้นจากผู้ใช้งานเอกสารอย่างถูกต้องและเหมาะสม

งานทางด้านการจัดหมวดหมู่เอกสารเป็นกลุ่มตามเนื้อหานั้น ได้รับความสนใจและมีการนำเสนอเทคนิควิธีการจัดกลุ่มหรือหมวดหมู่ด้วยการเรียนรู้ด้วยคอมพิวเตอร์หลายหลายวิธี ซึ่งเทคนิควิธีการเหล่านี้สามารถนำมาประยุกต์ใช้กับข้อมูลหลายประเภท ทั้งข้อความ รูปภาพและเสียง โดยการจัดแบ่งเอกสารนั้นสามารถแบ่งได้ 2 ลักษณะ ได้แก่ การจัดกลุ่ม (Clustering) และการจำแนกหมวดหมู่ (Classification หรือ Categorization) การจัดกลุ่มเอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสารโดยไม่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน ซึ่งจะเป็นการแบ่งกลุ่มตามลักษณะของเอกสาร โดยเอกสารที่มีลักษณะเหมือนกันจะอยู่ด้วยกัน ส่วนการจำแนกหมวดหมู่เอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสารโดยที่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน โดยจะเปรียบเทียบเอกสารกับต้นแบบในแต่ละหมวดหมู่เอกสารจะถูกจัดอยู่ในหมวดหมู่ที่ต้นแบบมีลักษณะคล้ายกับตัวเองมากที่สุด โดยใช้ใจความสำคัญของเอกสาร โดยใช้วิธีการทางด้านการเรียนรู้ด้วยคอมพิวเตอร์ (Machine Learning) ซึ่งเป็นการสร้างตัวจำแนกหมวดหมู่เอกสารอัตโนมัติ (Automatic Text Classifier) ด้วยเครื่องคอมพิวเตอร์ โดยการเรียนรู้ด้วยวิธีการเชิงอุปนัยจากชุดของเอกสารที่ได้จำแนกประเภทไว้ก่อนและ ลักษณะเฉพาะของหมวดหมู่ที่เกี่ยวข้อง ข้อดีของวิธีการทางด้านการเรียนรู้ด้วยคอมพิวเตอร์คือ ความถูกต้องของผลลัพธ์ที่ใกล้เคียงกับผลการจำแนกหมวดหมู่ของเอกสารที่ทำโดยมนุษย์และสามารถประหยัดแรงงานมนุษย์เป็นอย่างมาก เพราะไม่ต้องใช้มนุษย์ในการจำแนกประเภทเอกสารหรือปรับเปลี่ยนหมวดหมู่ของเอกสาร [1-2] จากความสำคัญข้างต้น ในบทความนี้มีวัตถุประสงค์ที่จะนำเสนอขั้นตอนวิธีการสร้างแบบจำลองการจำแนกเอกสารภาษาไทยอัตโนมัติ เพื่อแก้ปัญหาดังกล่าว โดยมีรายละเอียดหัวข้อในด้านต่างๆ ดังนี้

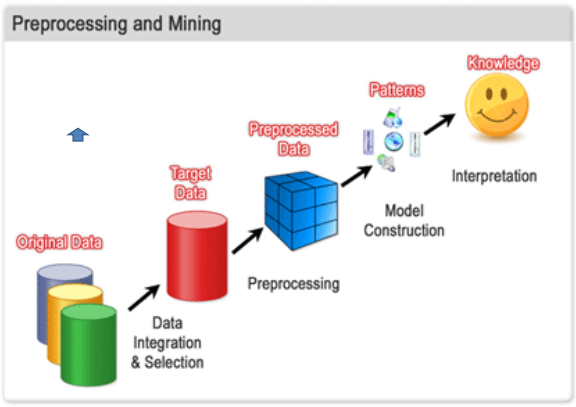
2. เหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล (Data Mining) หรืออาจจะเรียกว่าการค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery in Databases - KDD) เป็นเทคนิคเพื่อค้นหาภาพแบบ (Pattern) ของจากข้อมูลจำนวนมหาศาลโดยอัตโนมัติ จัดเป็นขบวนการของการดึงเอาความรู้ออกมาจากข้อมูลขนาดใหญ่ โดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่อง และการรู้จำแบบ หรือในอีกนิยามหนึ่ง การทำเหมืองข้อมูล คือกระบวนการที่กระทำกับข้อมูลจำนวนมาก เพื่อค้นหาภาพแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์ [3]

Data Mining ตามศัพท์ที่ราชบัณฑิตยสถานกำหนดไว้หมายถึง การสกัดหรือวิเคราะห์ ค้นหาข้อมูลที่ต้องการจากข้อมูลจำนวนมากได้หรือกล่าวอีกนัยหนึ่งคือ ชุดซอฟต์แวร์ (Software) วิเคราะห์ข้อมูลที่ได้ถูกออกแบบมาเพื่อระบบสนับสนุนความต้องการของผู้ใช้ในการค้นหาข้อมูลที่ต้องการจากข้อมูลจำนวนมากได้

ส่วนการนำเหมืองข้อมูลมาประยุกต์ในการสร้างแบบจำลองเอกสารนั้น ขั้นแรกจะต้องทำการเลือกเทคนิคที่เหมาะสมกับภาพแบบชุดข้อมูลเอกสารที่จะทดลอง โดยพิจารณาจากปัญหาของเอกสาร เช่น ภาษาที่ใช้ ประเภทของ

เอกสาร จำนวนคุณลักษณะจำนวนเอกสาร เป็นต้น หลังจากได้เทคนิคที่เหมาะสมกับชุดเอกสารแล้ว จะทำการสอน (Train) ให้แบบจำลองเรียนรู้ลักษณะของชุดข้อมูลเอกสารว่า ชุดข้อมูลทั้งหมดมีความสัมพันธ์กันอย่างไร และทิศทางในการจำแนกอย่างไร โดยในการสอนให้แบบจำลองเรียนรู้นั้น จำเป็นต้องมีการกำหนดพารามิเตอร์ (Parameter) หรือค่าตัวแปรต่างๆให้เหมาะสมกับชุดเอกสารที่นำมาเรียนรู้ ซึ่งในการพิจารณาค่าพารามิเตอร์นั้นขึ้นอยู่กับเทคนิคที่เลือกใช้ประสบการณ์ในการวิเคราะห์ จากนั้นจึงนำแบบจำลองที่ได้ไปทดสอบหาค่าความผิดพลาดของแบบจำลอง โดยการนำข้อมูลเอกสารจริงที่เตรียมไว้สำหรับการทดสอบมาป้อนลงในแบบจำลองแล้วดูผลของการทำนายที่ได้



รูปที่ 1 ขั้นตอนกระบวนการทำเหมืองข้อมูล

3. ขั้นตอนการจำแนกเอกสาร (Text Categorization)

จากการศึกษางานวิจัยที่เกี่ยวข้องกับการจำแนกเอกสารภาษาไทยอัตโนมัติ พบว่ามีขั้นตอนในการทำงานหลักในการทำงานดังต่อไปนี้ ในขั้นตอนแรกจะทำการสกัดคุณลักษณะด้วยการตัดคำเพื่อให้ได้คุณลักษณะจากเอกสารออกมา จากนั้นทำการกำจัดคำหยุดและทำรากศัพท์จากฐานข้อมูลภาษาไทยที่กำหนดขึ้น หลังจากนั้นทำการให้ค่าน้ำหนักดัชนีของคำในเอกสาร (Term weighting) แล้วทำการลดขนาดคุณลักษณะเพื่อมิติของเอกสารลง จากนั้นส่งเข้าเครื่องจักรการเรียนรู้แบบมีผู้สอน (Supervised Learning) แล้วทำการทดสอบเปรียบเทียบประสิทธิภาพด้านความถูกต้อง(Accuracy) ความแม่นยำ (Precision) การระลึก (Recall) [4-5] ดังสมการ

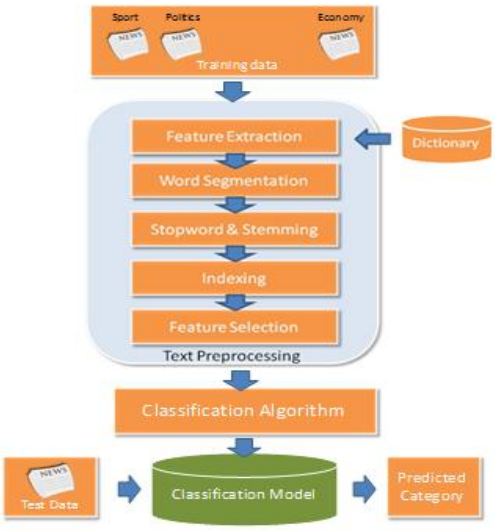
$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (3)$$

โดยใช้วิธีการประเมินความสามารถของแบบจำลอง โดยวัดที่ประสิทธิผลของการจำแนกหมวดหมู่ตามแนวคิดทางด้านการค้นคืนสารสนเทศคือการวัดค่า F1-measure ดังสมการ

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$



รูปที่ 2 แบบจำลองการจำแนกเอกสารภาษาไทย

4. การสกัดคุณลักษณะ (Feature Extraction)

เนื่องจากคอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ของเอกสารซึ่งเป็นภาษาธรรมชาติโดยตรงได้ ดังนั้นจึงต้องแปลงเอกสารให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถใช้ในการเรียนรู้ได้ซึ่งขั้นตอนนี้เราเรียกว่าการสกัดคุณลักษณะ วัตถุประสงค์ที่สำคัญของขั้นตอนการสกัดคุณลักษณะเอกสารคือการดึงคุณลักษณะ (Feature) ของเอกสารออกมา ซึ่งการดึงคุณลักษณะออกมานั้น ก่อนอื่นเราต้องการกำหนด

ก่อนว่าจะใช้อะไรเป็นตัวแทนคุณลักษณะของเอกสาร และใช้ค่าใดแทนคุณลักษณะเอกสารนั้น จากการสำรวจงานที่ผ่านมาทั้งในประเทศและต่างประเทศพบว่า สามารถแทนคุณลักษณะด้วยคำเดี่ยว พยางค์ วลี กลุ่มของคำ ประโยค เป็นตัวแทนคุณลักษณะของเอกสาร [6-7]

ตัวแทนคุณลักษณะของเอกสารที่นิยมใช้ในการจัดหมวดหมู่เอกสารประเภทข้อความคือ ถุงคำ (Bag of words) ซึ่งจะเก็บอยู่ในรูปแบบของเวกเตอร์ โดยองค์ประกอบของเวกเตอร์อาจจะแทนด้วยคุณลักษณะของค่าความจริง (Boolean) แทนด้วยค่าความถี่ของคำ(Term frequency) หรือแทนค่าน้ำหนักตามหลักการค้นคืนสารสนเทศ ด้วยวิธีนับค่าความถี่ของคำ ร่วมกับความถี่เอกสารผกผัน (Term Frequency-Inverse Document Frequency) ตลอดจนการค่าน้ำหนักของชุดลำดับคำ (N-Gram) หรือค่าน้ำหนักของวลีในเอกสาร มาใช้เป็นตัวแทนคุณลักษณะเอกสาร ดังนั้นขั้นตอนนี้จะเกี่ยวข้องกับวิธีการตัดคำ เช่น ตัดคำเดี่ยว การตัดคำโดยใช้ N-Gram การตัดคำเดียวร่วมกับชุดของคำ Bi-gram เป็นต้น นอกจากการตัดคำแล้ว ยังรวมถึงขั้นตอนในการลดขนาดเอกสารลงอีกด้วย วิธีที่ใช้ได้แก่ การนำคำที่ไม่มีนัยสำคัญออก และการวิธีทำรากศัพท์ ด้วยเช่นกัน [6-7]

5. การตัดคำ (Word Segmentation)

การประมวลผลจำแนกหมวดหมู่เอกสารภาษาไทยได้อย่างมีประสิทธิภาพนั้น ในส่วนของการประมวลผลเบื้องต้นคือการสกัดคุณลักษณะของเอกสารภาษาไทย ซึ่งขั้นตอนดังกล่าวมีความจำเป็นต้อง การตัดคำในภาษาไทย (Word Segmentation) แต่พบปัญหาเนื่องจากภาษาไทยไม่มีสัญลักษณ์ที่สามารถบ่งบอกถึงขอบเขตของคำ ที่เรียงต่อเนื่องกันทั้งประโยคเหมือนกับภาษาอังกฤษที่ใช้ช่องว่าง (Space) คั่นระหว่างขอบเขตของคำ ซึ่งเป็นอุปสรรคอย่างหนึ่งที่เพื่อตัดสายอักขระภาษาไทยออกเป็นคำๆ ได้อย่างถูกต้อง จึงได้มีผู้คิดค้นพัฒนาวิธีการตัดคำภาษาไทย (Thai Word Segmentation) ซึ่ง โดยในปัจจุบันยังมีการพัฒนาการตัดคำภาษาไทยอยู่เนืองๆ ซึ่งสามารถแบ่งตามลักษณะฐานข้อมูลที่นำมาใช้ในการตัดคำ ออกเป็น 3 กลุ่มหลัก คือ หลักการตัดคำโดยใช้กฎ (Rule-Based Approach) หลักการตัดคำโดยใช้พจนานุกรม (Dictionary-Based Approach) และหลักการตัดคำโดยใช้คลังข้อมูล (Corpus-Based Approach) [8]

งานวิจัยที่ที่เกี่ยวข้อง ได้แก่ งานวิจัยของ ยืน และวิวรรธน์ [9] จัดเป็นงานวิจัยการแบ่งพยางค์ด้วยพจนานุกรม ซึ่งถือว่าเป็นงานวิจัยงานแรกของการตัดพยางค์ที่มีการนำพจนานุกรมเข้ามาใช้ โดยจะจัดเก็บพยางค์ต่างๆไว้ในพจนานุกรม และมีการนำกฎไวยากรณ์ต่างๆ จำนวน 18 กฎ เข้ามาช่วยในกรณีที่ไม่พบพยางค์ในพจนานุกรม หลักการทำงานของกระบวนวิธีการตัดพยางค์ด้วยพจนานุกรม คือ จะตรวจสอบสายอักขระ (String) ที่เข้ามาจากซ้ายไปขวากับพยางค์ที่เก็บไว้ในพจนานกรม ในกรณีที่ตรวจสอบแล้วปรากฏว่าพบพยางค์มากกว่า 1 พยางค์ในพจนานุกรม ก็ให้เลือกแบ่งพยางค์โดยเลือกพยางค์ที่ยาวที่สุด แล้วทำต่อไปเรื่อยๆ จนจบสายอักขระ แต่ถ้ากรณีที่เลือกพยางค์ที่ยาวที่สุดแล้วทำให้เกิดพยางค์ที่ไม่ปรากฏในพจนานุกรมก็ยอมให้มีการย้อนรอย (Back Tracking) กลับไปเลือกพยางค์ที่ยาวรองมาแทน ซึ่งวิธีการนี้จะเป็นที่รู้จักกันในชื่อ การตัดคำ แบบเลือกคำยาวที่สุด (Longest Matching)

งานวิจัยของ วิรัช [10] เป็นงานวิจัยที่ได้พัฒนาการตัดคำโดยเรียกว่า การตัดคำโดยเลือกแบบเหมือนมากที่สุด (Maximal Matching) ซึ่งขั้นตอนวิธีนี้ จะสามารถแก้ไขความบกพร่องของการตัดคำแบบเลือกคำยาวที่สุดได้ โดยจุดบกพร่องที่กล่าวนี้คือ ขั้นตอนวิธีการตัดคำแบบเลือกคำยาวที่สุด ที่จะเลือกคำที่ยาวเกินไปตั้งแต่ครั้งแรก ทำให้ข้อความที่ตามมาเกิดข้อผิดพลาดได้ เช่นประโยค “ไปหามเหสี” จะตัดคำได้เป็น “ไป หาม เห สี” โดยที่ถูกต้องจะต้องตัดเป็น “ไป หา มเหสี”

งานวิจัยของ จันทิมา [11] นำเสนอระบบการจัดกลุ่มเอกสารข้อความอัตโนมัติด้วยซัพพอร์ตเวกเตอร์แมชชีน นำเสนอการจัดกลุ่มเอกสารข้อความภาษาไทยแบบอัตโนมัติ บนพื้นฐานของการตัดคำด้วยพจนานุกรม (Dictionary-Based Approach) ด้วยวิธีการตัดคำโดยใช้พจนานุกรมโดยการเทียบคำที่ยาวที่สุด (Longest Matching) และเรียนรู้ด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์ แมชชีน ผลลัพธ์จะการตัดคำด้วยวิธีการเทียบคำที่ยาวที่สุด เมื่อทำการเรียนรู้และจำแนกหมวดหมู่เอกสารภาษาไทย โดยทดสอบประสิทธิภาพของระบบด้วยการวัดค่า F-Measure

พบว่าให้ความถูกต้องในการจำแนกกลุ่มเอกสารอยู่ระหว่างร้อยละ 77.46% - 84.70% ซึ่งอยู่ในระดับดีมาก

จากงานวิจัยดังกล่าวพบว่า การตัดคำภาษาไทยมีการพัฒนาขึ้นด้วยกันหลากหลายวิธีนั้น แต่ละวิธีก็มีข้อดีและข้อเด่นของตนเอง และให้ผลลัพธ์ที่แตกต่างกันในด้านของความถูกต้อง ความรวดเร็วในการทำงาน และปริมาณการใช้ทรัพยากรแตกต่างกัน จากการศึกษาพบว่าวิธีการตัดคำที่เหมาะสมกับการจำแนกเอกสาร ได้แก่การตัดคำโดยใช้พจนานุกรมโดยการเทียบคำที่ยาวที่สุด (Longest Matching) จะส่งผลให้แบบจำลองการจำแนกเอกสารมีประสิทธิภาพมากที่สุด [2,5,11]

6. การกำจัดคำหยุด (Stop-Word List Removal)

เป็นการนำคำที่ไม่มีนัยสำคัญออกโดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลง เป็นคำที่ไม่มีนัยสำคัญในที่นี้หมายถึงคำที่ใช้กันโดยทั่วไปไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง คำหยุดมักเป็นคำที่ปรากฏขึ้นบ่อยครั้งในเอกสาร และปรากฏในเอกสารเกือบทุกฉบับ จึงถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการค้นคืนหรือการจำแนกหมวดหมู่ ดังนั้นการกำจัดคำหยุดจึงเป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี เพื่อกำจัดคุณลักษณะที่ไม่เป็นประโยชน์ และลดขนาดของดัชนีลง ซึ่งจะช่วยประหยัดทั้งพื้นที่และเวลาในการประมวลผล [7-8]

ดังงานวิจัยของ จิเรขา [12]ได้พัฒนาระบบจัดหมวดหมู่เอกสารภาษาไทยบนระบบเครือข่ายโดยใช้โครงข่ายประสาทเทียม สกัดคุณลักษณะด้วยวิธีการตัดคำแบบอิงพจนานุกรม ที่มีความสามารถในการช่วยลดคำที่ไม่รู้จักของเอกสารเนื่องจากสามารถเพิ่มคำที่เกี่ยวข้องกับหัวข้อที่ผู้ใช้ต้องการได้ ทำให้ระบบตัดคำลงในพจนานุกรมได้เพื่อให้ผลจากการตัดคำตรงตามความต้องการ และทำการนำคำที่ไม่นัยสำคัญออกจาก (Stopword list) ของฐานข้อมูลคำหยุดจากระบบเครือข่ายอินเทอร์เน็ตในหน้าหนังสือออนไลน์ของประเทศไทย เมื่อนำเอาฐานข้อมูลคำหยุดมาประยุกต์กับแบบจำลองการจำแนกเอกสาร ทำให้สามารถกำจัดคำที่ไม่มีนัยสำคัญของเอกสารออกได้เป็นจำนวนมาก เนื่องจากลักษณะการใช้ภาษาในงานเอกสารนั้น มักมีการใช้คำหยุดประเภทคำบุพบท คำสรรพนาม คำสันธาน อยู่เป็นจำนวนมากซึ่งไม่ส่งผลในการจำแนกเอกสาร

7. การหารากศัพท์ (Stemming)

เป็นการหารูปเดิมของคำ หรือหาคำที่มีความหมายคล้ายกัน เพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์เป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี ทำให้สามารถลดขนาดของดัชนีลงและเพิ่มประสิทธิภาพในการค้นคืนหรือการจำแนกหมวดหมู่ การหารากศัพท์ของคำภาษาอังกฤษมีขั้นตอนวิธีที่แน่นอนซึ่งสามารถเขียนเป็นอัลกอริทึมในการสกัดคำอุปสรรคและคำปัจจัยได้โดยไม่ต้องเทียบกับรายการคำศัพท์หรือคลังคำ เนื่องมาจากไวยากรณ์ของภาษาอังกฤษมีกฎเกณฑ์ที่แน่นอน ไม่ค่อยมีข้อยกเว้นมากนัก จึงมีความซับซ้อนน้อย ดังนั้นจึงมีผู้สร้างอัลกอริทึมสำหรับการหารากศัพท์ไว้หลายแบบ เช่น Porter Algorithm [6-7] เป็นต้น

8. การสร้างดัชนีเอกสาร (Document Indexing)

เนื่องจากคอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ของเอกสารซึ่งเป็นภาษาธรรมชาติโดยตรงได้ ดังนั้นจึงต้องแปลงเอกสารให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถใช้ในการเรียนรู้ได้ ขั้นตอนในการแปลงเอกสาร เรียกว่า การทำดัชนี (Indexing) เพื่อสร้างตัวแทนเนื้อหาของเอกสาร (Document Representation) สำหรับใช้ในกระบวนการเรียนรู้ ลักษณะของตัวแทนเอกสารขึ้นอยู่กับสิ่งที่ต้องการพิจารณาว่า ต้องการพิจารณาเฉพาะความหมายของคำ หรือต้องการพิจารณาความหมายตามกฎของภาษา ซึ่งจะสนใจตำแหน่งของคำที่อยู่ในประโยค การจำแนกหมวดหมู่ด้วยวิธีการทางด้านการเรียนรู้ด้วยคอมพิวเตอร์ นิยมใช้ลักษณะของตัวแทนเอกสารที่สนใจความหมายของคำ โดยไม่สนใจตำแหน่งของคำ ซึ่งตัวแทนเอกสารมักจะอยู่ในรูปของเวกเตอร์ของน้ำหนักคำ วัตถุประสงค์ที่สำคัญในส่วนนี้คือการสร้างดัชนี ซึ่งก็คือการคำนวณหาค่าที่จะมาใช้เป็นค่าคุณลักษณะของเอกสาร หรืออาจจะเรียกได้ว่าการหาค่าน้ำหนัก (Term weighting) ซึ่งคุณลักษณะอาจหมายถึง คำเดี่ยว (Single Word) รากศัพท์ (Stems) วลี (Phrase) ชุดลำดับคำ

(N-Gram) ประโยค (Sentences) หรือสิ่งอื่น แต่โดยมากนิยมใช้ในรูปแบบของคำเดี่ยว [4-7]

	คำศัพท์	เอกสาร$_1$	เอกสาร$_2$	เอกสาร$_3$	เอกสาร$_4$	เอกสาร$_N$
1	ศาล ปกครอง	3	0	0	0	1
2	สูงสุด	2	3	1	2	1
3	เลื่อน	2	2	2	1	1
4	กทช.	4	0	0	0	1
..						
M	อุทธรณ์	4	0	0	1	0

รูปที่ 3 เวกเตอร์ตัวแทนเอกสาร

โดยทั่วไปที่นิยมใช้กัน จะเริ่มจากการสร้างเวกเตอร์ตัวแทนเอกสาร จากนั้นจะสร้างเมตริกซ์ของกลุ่มเอกสารขึ้นจากเวกเตอร์เอกสารทั้งหมดในกลุ่ม ซึ่งคำนวณค่าน้ำหนักให้ กับดัชนีส่วนใหญ่ใช้วิธีการ Tfidf-weighting (Term frequency–inverse document frequency)

$$tfidf_{ik} = 1 + \log(tf_{ik}) \times \log\left(\frac{N}{n_i}\right) \quad (5)$$

ดังงานวิจัยของ ชูชาติ [13] นำเสนอการให้ค่าน้ำหนักดัชนี ด้วยวิธี tfidf-weighting โดยทดลองกับกลุ่มตัวอย่างหนังสือพิมพ์ไทยรัฐจากเว็บไซต์ โดยใช้การตัดคำด้วยคำเดี่ยว (Single word) ผลการทดลองพบว่าเมื่อใช้อัลกอริทึม Support Vector Machines ให้ประสิทธิภาพออกมาสูงสุด รองลงมาเป็น Naïve Baye และ Decision Tree ตามลำดับ การให้ค่าดัชนีในแทนคำในแบบจำลองการจำแนกเอกสารนั้น มีการให้ค่าน้ำหนักหลากหลายวิธี แต่ละวิธีก็มีข้อดีและให้ ผลลัพธ์ที่แตกต่างกัน จากการศึกษาพบว่าวิธีการให้ค่าน้ำหนัก แบบ tfidf-weighting ซึ่งเป็นวิธีที่นิยมในการนำมาใช้ในการสร้างแบบจำลองเอกสาร เนื่องจากมีประสิทธิภาพในการจำแนกอยู่ในระดับดี และให้ค่าความแม่นยำในจำแนกหมวดหมู่สูงและโดยวิธีการคำนวณไม่ซับซ้อน ใช้หน่วยประมวลผลไม่มากนัก

9. การเลือกคุณลักษณะ (Feature Selection)

จากที่กล่าวมาจะพบว่าเอกสารนั้นมีแนวโน้มในการที่จะเพิ่มปริมาณสูงขึ้นทุกวัน ทำให้เอกสารมีจำนวนคุณลักษณะมากขึ้น ทำให้วิธีเบื้องต้นในการลดขนาดเอกสารคือการใช้การนำคำที่ไม่มีนัยสำคัญออก กับการทำรากศัพท์แล้วยังไม่เพียงพอ ซึ่งจำนวนคุณลักษณะมีผลต่อประสิทธิภาพของการจำแนกหมวดหมู่เอกสาร เนื่องจากอัลกอริทึมที่ใช้ในการเรียนรู้เพื่อสร้างตัวจำแนกหมวดหมู่ โดยทั่วไปไม่สามารถรองรับการทำงานกับจำนวนคุณลักษณะของเอกสารที่สูงมากได้ดี การลดขนาดเอกสารจึงเป็นขั้นตอนหนึ่งที่จะต้องทำก่อน การสร้างตัวจำแนกเอกสารแต่การลดขนาดของเอกสารต้องพิจารณาด้วยความระมัดระวัง เนื่องจากมีความเสี่ยงในการที่จะกำจัดคุณลักษณะที่สำคัญต่อการจำแนกหมวดหมู่ออกไป จากการศึกษาพบว่าวิธีการลดคุณลักษะเอกสารประกอบด้วยการสร้างคุณลักษณะใหม่จากคุณลักษณะเดิม อาจจะนำคุณลักษณะพื้นฐานเหล่านี้มารวมกันเพื่อให้เป็นคุณลักษณะใหม่ที่มีระดับสูงขึ้นการเลือกคุณลักษณะที่นิยมใช้ [2,5-7] ได้แก่ ค่าความถี่เอกสาร (DF: Document Frequency) ค่าสารสนเทศ (IG: Information Gain) ค่าสถิติไคสแควร์ (Chisquare) เป็นต้น

ดังงานวิจัยของ นิเวศ [2,18] และ ชูชาติ [13] ได้นำเสนอวิธีการลดคุณลักษณะเอกสารร่วมกับอัลกอริทึมเครื่องจักรการเรียนรู้ เพื่อลดระยะเวลาและทรัพยากรของระบบในการประมวลผล โดยทำการศึกษาการลดคุณลักษณะด้วยวิธี Information Gain วิธีChi-square และวิธี Document Frequency จากการประยุกต์ใช้วิธีการลดคุณลักษณะกับแบบจำลองการจำแนกเอกสารนั้น มีการหลากหลายวิธี แต่ที่ได้รับความนิยมสูง เนื่องจากมีประสิทธิภาพดี และใช้เวลาในการประมวลผลไม่นานนัก ได้แก่ วิธีการ Information Gain Ranking จากการศึกษาพบว่า การลดขนาดคุณลักษณะของกลุ่มเอกสารที่ใช้วิธีการดังกล่าว สามารถลดได้มากถึง 80-90 % โดยการลดลงของคุณลักษณะดังกล่าวนั้น ไม่ส่งผลให้ประสิทธิภาพในการจัดจำแนกเอกสารลดลงแต่อย่างใด

10. อัลกอริทึมการจำแนกเอกสาร (Classification Algorithm)

จัดเป็นกระบวนการที่ใช้ในการหาแบบจำลองของชุดข้อมูลที่มีความใกล้เคียงกันหรือเหมือนกันมากที่สุด เพื่อใช้ในการทำนายชุดข้อมูลใหม่ว่าอยู่ในประเภทใดของชุดข้อมูลใดที่ได้ทำการแบ่งไว้แล้ว ซึ่งชุดข้อมูลที่แบ่งไว้เกิดจากการเรียนรู้จากชุดข้อมูลที่มีอยู่แล้ว (Training Data) แบบจำลองที่เกิดจากการเรียนรู้ สามารถแสดงได้หลายภาพแบบ เช่น กฎการแบ่ง (Classification Rules, IF-THEN) การคำนวณแบบต้นไม้วิเคราะห์ (Decision Tree) การใช้สูตรทางคณิตศาสตร์ (Mathematical Formula) หรือโครงข่ายใยประสาทเทียม(Neural Network) เป็นต้น ในส่วนของการทำต้นไม้วิเคราะห์ จะแสดงออกมาในลักษณะของแผนภูมิโครงสร้างต้นไม้ ซึ่งก้านของต้นไม้จะแสดงถึงความรู้ที่ได้และใบไม้จะแสดงถึงประเภทชุดข้อมูลที่ถูกแบ่งออกมา แผนภูมิต้นไม้สามารถแปลงเป็นกฎการแบ่งได้ง่ายเพราะลักษณะของแผนภูมิสามารถเข้าใจได้ง่าย ในส่วนของโครงข่ายใยประสาทเทียมนั้น จะแสดงในลักษณะของการเชื่อมต่อระหว่างหน่วยที่เกิดขึ้น ตัวอย่างเทคนิคของการจัดหมวดหมู่ ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) [17-19] เนอีฟเบย์ (Naïve-Bayes) การคำนวณแบบพันธุกรรม (Genetic Algorithm) และโครงข่ายใยประสาทเทียม (Neural Network) เป็นต้น [12-16] ตัวอย่างงานวิจัยแนวนี้คือ Joachim [17] ใช้หลักการของ TFIDF (Term Frequency and Inverse Document Frequency) โดยผสมผสานแนวคิดของความน่าจะเป็นของเบย์สอย่างง่ายเข้าไปด้วย ซึ่งให้ผลการทดลองในการจำแนกค่อนข้างดีกว่าการจำแนกด้วยวิธีของเบย์ส เพียงอย่างเดียว และ Joachim ได้เปรียบเทียบการใช้งาน 5 อัลกอริทึมในการจัดหมวดหมู่ ได้แก่ Naïve-Bayes Rocchio C4.5 k-NN และ SVM บนข้อมูลชุด Reuters-21578 ผลลัพธ์ที่ได้ออกมาโดยดูจากค่า Break-even ปรากฏว่า Support Vector Machine ได้ผลลัพธ์ในการจำแนกออกมาดีที่สุด สอดคล้องกับงานวิจัยของ นิเวศ [2,18] และชูชาติ [13] ที่ทำการทดสอบประสิทธิภาพในการจำแนกเอกสาร พบว่าอัลกอริทึม Support Vector Machine มีประสิทธิภาพสูงที่สุด รองลงมาเป็นอัลกอริทึม Naïve-Bayes และอัลกอริทึม Decision Tree ตามลำดับ

จากการศึกษาพบว่าอัลกอริทึมที่นิยมในการสร้างแบบจำลองการจำแนกเอกสารนั้น จะมีทั้งการจำแนกประเภทโดยให้ทฤษฎีความน่าจะเป็น ต้นไม้การตัดสินใจ ระนาบการตัดสินใจ และอื่นๆ จากผลการทดลองพบว่าอัลกอริทึมที่มีประสิทธิภาพในการจำแนกเอกสารสูงและใช้เวลาในการประมวลผลไม่มากนักได้แก่ Support Vector Machine

11. บทสรุป

บทความวิชาการนี้ได้ทำการรวบรวมเอาทฤษฎีและงานวิจัยต่างๆที่เกี่ยวข้อง ในการสร้างแบบจำลองการจำแนกเอกสาร มาสรุปและแจกแจงแสดงขั้นตอนวิธีการแบบละเอียด พร้อมทั้งยกงานวิจัยประกอบเป็นลำดับขั้น ซึ่งจะสามารถช่วยให้ผู้อ่านทำเข้าใจขั้นตอนวิธีการสร้างแบบจำลองการจำแนกเอกสารแบบอัตโนมัติได้โดยง่าย โดยแบบแบบจำลองการจำแนกหมู่เอกสารแบบอัตโนมัตินั้น ถูกใช้เพื่อแก้ปัญหาการจำแนกประเภทของเอกสารที่ทำโดยมนุษย์ เนื่องจากเอกสารอิเล็กทรอนิกส์มีปริมาณมากขึ้นทุกวัน ความต้องการของทรัพยากรบุคคลในการจำแนกเอกสารเหล่านี้ต้องมีมากตามไปด้วยเช่นกัน จากเหตุผลดังกล่าวการแบบจำลองการจำแนกเอกสารอัตโนมัติด้วยคอมพิวเตอร์ จึงถูกพัฒนาขึ้นเพื่อประโยชน์ในการจำแนกเอกสารและจัดการกับข้อมูลแบบอัตโนมัติ โดยแบบจำลองที่นำเสนอนี้สามารถนำไปประยุกต์ในการสร้างระบบคัดกรองเอกสาร (Document Filtering) ระบบการค้นคืนเอกสารอัตโนมัติ (Automatic Indexing for IR System) ระบบการจัดหมวดหมู่ของเว็บเพจอัตโนมัติ (Web Page Classification) เป็นต้น

12. เอกสารอ้างอิง

[1] Sebastiani, Fabrizio, “Machine Learning in Automated Text Categorization”, ACM Computing Surveys, 2002.

[2] N. Chirawichitchai, P. Sanguansat, P. Meesad, “An Experimental Study on Feature Reduction Techniques and Classification Algorithms of Thai Documents”, Journal of Science Ladkrabang, 2009. (in Thai)

[3] Han and Kamber, “Data Mining Concepts and Techniques” , SanFrancisco, Morgan Kaufmann Publishers, 2006.

[4] N. Chirawichitchai, P. Sanguansat and P. Meesad "A Comparative Study on Term Weight Techniques for Thai Document Categorization". Journal of Science Ladkrabang, 2010. (in Thai)

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

[6] Aas, and Eikvil, “Text Categorization: a Survey”, Report No. 941, Norwegian Computing Center, 1999.

[7] V. Incham, "Automatic Thai document categorization system using SVM with language processing", Department of Computer Engineering,Kasetsart University, 2548. (in Thai)

[8] V. Sornlertlamvanich , "Word Segmentation for Thai in Machine Translation System", Machine Translation, NECTEC, 2536. (in Thai)

[9] Y. Poovarawan, "Thai Syllable Separater by Dictionary" , Electrical Engineering Conference, 2529. (in Thai)

[10] V. Sornlertlamvanich, "Word Segmentation for Thai in Machine Translation System". Machine Translation, National Electronics and Computer Technology Center, Bangkok. pp. 50-56, 1993. (in Thai)

[11] J. Polpinij. "Automatic document classification by support vector machines", Mahasarakham University, 2548. (in Thai)

[12] J. Thaisungkhom, "The document classification system based on the neural network", King Mongkut’s Institute of Technology North Bangkok , 2549 (in Thai)

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

[14] N. Sura, "Automatic Thai Text Categorization Using FPTC Algorithm", Department of Computer Science, King Mongkut’s Institute of Technology North Bangkok, 2549. (in Thai)

[15] W. Cohen, and Singer, “Context-Sensitive Learning Methods for Text Categorization” , Proceedings of SIGIR-96, 19th ACM International Conference On Research and Development in Information Retrieval, 1998.

[16] Lertnattee and Theeramunkong, “Text Classification for Thai Medicinal Web Pages” , PAKDD, 2007.

[17] Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”, LS8-Report 23, Universität Dortmund, LS VIII-Report, 1997.

[18] N.Chirawichitchai " Sentiment Classification Using Data Reduction and Term Indexing Techniques", Ladkrabang Engineering Journal, 2011, pp.25-30. (in Thai)

[19] S.chunwiphate "Data Classification Problem Solving using Support Vector Machine Approachand Its Application", The Journal of Industrial Technology, 7, 2011, pp. 50-55. (in Thai)