

การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะแบบการกรองและการควบรวม ของการทำเหมืองข้อความเพื่อการจำแนกข้อความ

วาที น้อยเพียร^{1,2*} และ พยุง มีสัจ³

บทคัดย่อ

ปัญหาหนึ่งของการทำเหมืองข้อความคือข้อมูลมีปริมาณมาก นักวิจัยจำนวนมากใช้เทคนิคการคัดเลือกคุณลักษณะเพื่อได้ค่าที่เหมาะสมในการแทนเอกสารและเพิ่มประสิทธิภาพในการจำแนกเอกสารให้มีค่าความถูกต้องมากขึ้น เทคนิคที่ใช้แบ่งเป็น 2 วิธี ได้แก่ การกรองและการควบรวม โดยเทคนิคการควบรวมสามารถใช้เทคนิคการทำเหมืองข้อความร่วมกับการค้นหาข้อมูล ในงานวิจัยนี้ได้ทำการเปรียบเทียบการคัดเลือกคุณลักษณะแบบการกรอง โดยเลือกใช้อินฟอรมชันแกน เคนเรโซ และไคสแควร์ วิธีคัดเลือกแบบไคสแควร์ให้ผลดีที่สุดวัดประสิทธิภาพโดยรวม 92.2% และการควบรวมใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (SVM) ร่วมกับการค้นหาด้วยวิธีเชิงพันธุกรรม (SVMGA) และการค้นหาด้วยวิธีละโมบ (SVMGD) โดยวิธีคัดเลือกแบบ SVMGD ให้ผลดีที่สุดวัดประสิทธิภาพโดยรวม 94% ซึ่งการจำแนกข้อความทั้งสองวิธีใช้ขั้นตอนวิธีแบบซัพพอร์ตเวกเตอร์แมชชีนโดยใช้เคอร์เนลแบบเรเดียลเบสิสฟังก์ชัน (SVMR) เมื่อเปรียบเทียบประสิทธิภาพทั้งวิธีการกรองและการควบรวมสรุปได้ว่าประสิทธิภาพโดยรวมของการควบรวมมีค่ามากกว่าการกรอง 1.8% ซึ่งทำให้นักวิจัยสามารถนำเทคนิคของการควบรวมไปใช้เพิ่มประสิทธิภาพการจำแนกข้อความ

คำสำคัญ : การทำเหมืองข้อความ, การคัดเลือกคุณลักษณะแบบการกรอง, การคัดเลือกคุณลักษณะแบบการควบรวม, การค้นหาด้วยวิธีเชิงพันธุกรรม, การค้นหาด้วยวิธีละโมบ, ซัพพอร์ตเวกเตอร์แมชชีน

¹ สำนักคอมพิวเตอร์และเทคโนโลยีสารสนเทศ, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

² ภาควิชาคอมพิวเตอร์ศึกษา คณะครุศาสตร์อุตสาหกรรม, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

³ คณะเทคโนโลยีสารสนเทศ, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

* ผู้ติดต่อ, อีเมล: vtn@kmutnb.ac.th รับเมื่อ 27 พฤษภาคม 2556 ตอบรับเมื่อ 11 กันยายน 2556

A Comparison of Filter and Wrapper Approaches with Text Mining for Text Classification

Vatinee Nuipian^{1,2*} and Phayung Meesad³

Abstract

The main problem for text categorization is the highest dimensionality of feature space. Many researchers focus on instruction feature selection techniques to represent a document which in turn, increases the overall efficiency of a classification model. There are two general feature selection approaches: the Filter approach and the Wrapper approach. The Filter approach used Information Gain, Gain Ratio and Chi-square. The results showed that Chi-Square had highest performance with F-measure equaling 92.2%, the Wrapper approach used Support Vector Machine consisting of Genetic Algorithm (SVMGA) and Greedy (SVMGD). The results also found that Greedy (SVMGD) was the best algorithm with F-measure which equaled 94%. Both feature selection approaches employed Support Vector Machine with kernel Radial basis function as a classifier. When comparing the effectiveness of Filter approaches to Wrapper approaches, evaluated via F-measure shown that the value of Wrapper approaches were higher than that of Filter approaches at 1.8%. In conclusion, this technique enables researchers to increase the efficiency of a wrapper approach when implemented for information classification.

Keywords : Text Mining, Filter Approach, Wrapper Approach, Genetic Algorithm, Greedy, Support Vector Machine

¹ Institute of Computer and Information Technology, King Mongkut University of Technology North Bangkok.

² Department of Computer Education, Faculty of Technical Education, King Mongkut University of Technology North Bangkok.

³ Faculty of Information Technology, King Mongkut University of Technology North Bangkok.

* Corresponding author, E-mail: vtn@kmutnb.ac.th Received 27 May 2013, Accepted 11 September 2013

1. บทนำ

การจำแนกเอกสารเป็นวิธีหนึ่งที่ใช้มาช่วยจัดการกับกลุ่มเอกสารที่มีจำนวนมากขึ้น เช่น ข้อมูลในอินเทอร์เน็ต จึงทำให้การค้นคืนไม่ตรงตามความต้องการของผู้ใช้ทำให้ไม่สามารถสรุปความ ประมวลความหมายหรือหาความสัมพันธ์ของคำได้อย่างตรงประเด็น จึงมีนักวิจัยหลายท่านพยายามพัฒนาระบบการค้นคืนเชิงความหมาย โดยขั้นตอนแรกอาจใช้เทคนิคการทำเหมืองข้อความเข้ามาช่วยเพื่อให้คอมพิวเตอร์สามารถทำงานได้แบบอัตโนมัติ หรือกึ่งอัตโนมัติ ซึ่งผลลัพธ์ของเทอมที่ได้จากการคัดเลือกคุณลักษณะที่ดีสามารถนำมาสร้างคลาสตามความต้องการของการสร้างออนโทโลยีหรือระบบค้นคืนเชิงความหมายและการทำงานในส่วนนี้สามารถแบ่งเบาภาระของผู้เชี่ยวชาญได้เป็นอย่างดี การใช้เทคนิคการทำเหมืองข้อความจัดการกับเทอม เพื่อทำการทดสอบประสิทธิภาพของการจำแนกข้อความแต่ปัญหาหนึ่งที่พบเช่นข้อมูลมีมิติมาก เนื่องจากรายละเอียดของชุดข้อมูลที่ดีต้องมีความถูกต้อง น่าเชื่อถือ และครบถ้วน โดยเฉพาะเอกสารเป็นชุดข้อความที่มีเทอมไม่ซ้ำหรือคำที่เกิดขึ้นในเอกสารทั้งหมด สามารถเกิดขึ้นเป็นหลักล้านคำทำให้ข้อมูลมีมิติมากนักวิจัยจึงต้องประยุกต์ใช้อัลกอริทึมต่างๆ เพื่อการลดมิติข้อมูล เช่นการเลือกคำที่มีความถี่มากกว่าค่าที่กำหนด (Document Frequency) [1-2] เพื่อประหยัดทรัพยากรและใช้เวลาประมวลผลน้อย ดังนั้นนักวิจัยส่วนหนึ่งจึงใช้วิธีการคัดเลือกคุณลักษณะของข้อมูลโดยการทำให้อัตราส่วนข้อมูลเดิมมีขนาดลดลงและสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุด การคัดเลือกคุณลักษณะในปัจจุบันใช้แบบการกรอง (Filter Approach) และการรวบรวม (Wrapper

Approach) ซึ่งการกรองส่วนใหญ่เลือกใช้เทคนิคแบบอินฟอร์เมชันเกน (Information Gain: IG) เกนเรโซ (Gain Ratio : GR) ไคสแควร์ (Chi-square : χ^2) Haruechaiyasak และคณะ [1] และ Thongklin [2] ได้ใช้วิธีการคัดเลือกคุณลักษณะเป็นวิธีที่สำคัญสำหรับการเพิ่มประสิทธิภาพและการหาค่าความถูกต้องของการจัดกลุ่มข้อมูล หลังจากนั้นจึงวัดประสิทธิภาพของการจำแนกข้อความ ด้วยวิธีการที่หลากหลาย เช่น เคเนียร์สเต็นเบอร์ (K-Nearest Neighbor: KNN) [1], [3] เบย์ (Bayesian Network: BN) [1-4] และ ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) [1], [3] ส่วนวิธีการรวบรวม เป็นการนำอัลกอริทึมของการจำแนกข้อความร่วมกับการค้นหาข้อมูล เพื่อใช้ในการคัดเลือกคุณลักษณะข้อมูล [3, 10] ซึ่ง Saengsiri และคณะ [3] ได้นำวิธีค้นหาชุดของลักษณะ คือ การค้นหาด้วยวิธีเชิงพันธุกรรม (Genetic Algorithm) และการค้นหาด้วยวิธีละโมภ (Greedy Search) และทำการเปรียบเทียบวิธีการคัดเลือกข้อมูลที่ดีที่สุด

ดังนั้นในงานวิจัยนี้ ผู้วิจัยจึงทำการคัดเลือกคุณลักษณะข้อมูลที่ดีด้วยวิธีการเปรียบเทียบการคัดเลือกคุณลักษณะแบบการกรอง และการรวบรวมเพื่อเลือกใช้อัลกอริทึมที่ดีที่สุดสำหรับการจำแนกข้อความ และนำไปสร้างคลาสในส่วนของออนโทโลยี สำหรับการพัฒนาบบค้นคืนเชิงความหมายแบบกึ่งอัตโนมัติ

2. ทฤษฎีที่เกี่ยวข้อง

การทำเหมืองข้อความเริ่มจากการตัดคำเนื่องจากการค้นหาส่วนใหญ่ใช้หลักการค้นคืนจากคำสำคัญ ซึ่งใช้วิธีเทียบคำค้นกับเอกสาร (Matching) ที่มีอยู่ในฐานข้อมูล จึงใช้การตัดคำแบบคำเดี่ยว (Single Terms)

แต่ปัญหาหนึ่งที่พบคือข้อมูลมีมิติมากจึงต้องใช้ ขบวนการในการคัดเลือกคุณลักษณะ (Feature Selection) ที่เหมาะสม ประกอบด้วย 2 แบบ คือ การ กรอง และการควบรวม หลังจากนั้นจึงทำการจำแนก ข้อความด้วยวิธีเคเนียร์สเตนเบอร์ [4] เบย์ [1-4] และซัพพอร์ตเวกเตอร์แมชชีน [2]

2.1 การสกัดข้อความ (Text Extraction)

การสกัดข้อความคือเลือกสกัดเฉพาะส่วนที่ต้องการ เช่น ชื่อผู้แต่ง ชื่อเรื่อง คำสำคัญ ฯลฯ เนื่องจาก คอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ของเอกสาร ซึ่งเป็นภาษาธรรมชาติได้โดยตรง ดังนั้นจึงต้องแปลง เอกสารให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถใช้ในการ เรียนรู้ได้วัตถุประสงค์ที่สำคัญคือการดึงคุณลักษณะ ของเอกสารมาแสดง ซึ่งจากการสำรวจงานวิจัยที่ผ่านมา พบว่าสามารถแทนคุณลักษณะด้วยคำเดี่ยว พยางค์ วลี กลุ่มของคำ ประโยค เพื่อใช้เป็นตัวแทนของเอกสาร และใช้ค่าความถี่ของคำที่ปรากฏในเอกสารเป็นค่าของ คุณลักษณะ [1], [5], [6] โดยต้องผ่านกระบวนการหา รากศัพท์ และการกำจัดคำหยุด

2.1.1 การหารากศัพท์ (Stemming)

รูปแบบของคำที่ยังไม่เติมคำหน้า (Prefix) หรือคำ ท้าย (Suffix) เพื่อจัดคำหลายคำที่มีความหมายคล้ายคลึง กันเป็นคำประเภทเดียวกัน เช่น Compatible กับ Compatibility ตัด -ible กับ ibility ซึ่งรากศัพท์ (Stem) ที่เหลือคือคำเดียวกัน [1, 4]

2.1.2 การกำจัดคำหยุด (Stop Words)

คือการนำคำที่ไม่มีนัยสำคัญออกไป โดยไม่ทำให้ ความหมายของเอกสารเปลี่ยน เช่น he, and, in, or, all, again [1, 4]

2.2 การคัดเลือกคุณลักษณะ (Feature Selection)

การคัดเลือกคุณลักษณะของเอกสาร คือการนำ เอกสารจากระบบต่าง ๆ เช่น เอกสารข่าวจากเว็บไซต์ บทคัดย่อ มาแปลงเพื่อให้เอกสารอยู่ในรูปแบบเดียวกัน มีองค์ประกอบของเอกสารที่เหมือนกัน และแทน คุณลักษณะในเอกสารโดยใช้คำเดี่ยว พยางค์ วลี กลุ่มของคำ ประโยค เพื่อเป็นตัวแทนคุณลักษณะของ เอกสาร และทำการคำนวณหาค่าน้ำหนักของ คุณลักษณะ อาจใช้เทคนิคการวิเคราะห์ทางภาษาเข้ามา ช่วย (Language Analysis) เพื่อตัดคำที่ไม่จำเป็นออก หรือตัดคำตามสถิติของคลังประโยค เพื่อลดมิติของ ขนาดเอกสาร

เนื่องจากชุดของเอกสารเป็นชุดข้อความที่มีเทอมไม่ ซ้ำหรือคำที่เกิดขึ้นในเอกสารทั้งหมด โอกาสเกิดของคำ สามารถเกิดขึ้นเป็นหลักล้านคำทำให้ข้อมูลมีมิติมาก จึง ต้องประยุกต์ใช้การคัดเลือกคุณลักษณะที่ดีเพื่อนำมาหา ประสิทธิภาพ ลดเวลาในการประมวลผลและทรัพยากร [1, 3-4] การคัดเลือกคุณลักษณะแบ่งได้ 2 วิธี ได้แก่การ กรอง (Filter Approach) และการควบรวม (Wrapper Approach)

2.2.1 การกรอง

การกรอง คือการคัดเลือกคุณลักษณะซึ่ง Meesad และคณะ [7] ได้ทำการเปรียบเทียบการลดมิติข้อมูล หลาย ๆ แบบเพื่อเลือกใช้เทคนิคที่ดีที่สุด คือ ไคสแควร์ และ เกนเรโซ ส่วน Haruechaiyasak และคณะ [1] แนะนำอินฟอร์เมชันเกน เพื่อใช้ในการคัดเลือก คุณลักษณะที่ดีที่สุดของข้อมูลเพื่อใช้ในการจำแนก ข้อความ ประมวลผลได้รวดเร็วและใช้ทรัพยากรน้อย

1) อินฟอร์เมชันเกน (IG) คือการประเมินค่าเพื่อใช้ ในการแบ่งข้อมูลด้วยการคำนวณค่า Gain สำหรับแต่ละ

มิติข้อมูลถ้ามิติข้อมูลใดมีค่า Gain สูงสุด จะถูกเลือกให้เป็นกลุ่มย่อยที่มีอำนาจจำแนก ดังสมการที่ 1 แสดงการคำนวณค่า Entropy และคำนวณค่า Gain [8]

$$Entropy(p) = - \sum_{i=0}^{c-1} p(j|t) \log_2 p(j|t) \quad (1)$$

โดยที่ $\sum_i$ คือผลรวมของความน่าจะเป็นของค่า j ที่เกิดในคลาส t

$p(j|t)$ คือค่าความถี่ที่มีความสัมพันธ์ของกลุ่ม j กับ โหนด t

$$Gain = Entropy(p) - \left(\sum_{i=1}^k \frac{n_i}{n} Entropy(i) \right) \quad (2)$$

โดยที่ $Entropy(p)$ คือค่า Entropy ของตัว Root

$\sum_{i=1}^k \frac{n_i}{n} Entropy(i)$ คือค่า Entropy ในแต่ละโหนดย่อย

2) เกนเรโซ (GR) เป็นการประเมินความน่าเชื่อถือของมิติข้อมูลโดยการวัด Gain Ratio ในแต่ละคลาสการคำนวณ GR โดยใช้ค่า SplitINFO ในสมการที่ 3 และการคำนวณค่าการวัด Gain Ratio [9] ดังสมการที่ 4

$$SplitINFO = - \sum_{i=1}^k \frac{n_i}{n} \log_2 \frac{n_i}{n} \quad (3)$$

$$GainRatio = \frac{\Delta INFO}{SplitINFO} \quad (4)$$

3) ไคสแควร์ (χ^2) คือการประเมินค่าของ แอททริบิวต์ โดยคำนวณค่า Chi-Square ทางสถิติ [3] ดังแสดงใน สมการที่ 5

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (5)$$

2.2.2 การควบรวม (Wrapper Approach)

การควบรวม คือการนำอัลกอริทึมของการทำเหมืองข้อความร่วมกับการค้นหาข้อมูล มาช่วยในการคัดเลือกคุณลักษณะของข้อมูล จึงให้ผลลัพธ์ที่ดีกว่าแบบการกรอง แต่ใช้เวลาในการคำนวณมากกว่า ซึ่ง Saengsiri และคณะ [3] ได้แนะนำวิธีค้นหาชุดของคุณลักษณะ คือการค้นหาด้วยวิธีเชิงพันธุกรรม (Genetic Algorithm) และการค้นหาด้วยวิธีละโมภ (Greedy Search) โดยใช้วิธีการค้นแบบเดินหน้า (Forward Selection) และย้อนกลับ (Backward Selection) และทำการเปรียบเทียบวิธีการคัดเลือกข้อมูลด้วยวิธีการกรอง และวิธีการควบรวม ซึ่งวิธีการควบรวมให้ค่าความถูกต้องที่ดีกว่า สุคนธ์ทิพย์ [10] ได้ทำการเปรียบเทียบการคัดเลือกคุณลักษณะที่เหมาะสมของวิธีการกรองและการควบรวม ซึ่งพบว่าวิธีการ Hybrid Classification ที่ใช้การค้นหาด้วยวิธีเชิงพันธุกรรมร่วมกับ Wrapper โดยใช้อัลกอริทึม C4.5 ให้ค่าความถูกต้องสูงที่สุดและสามารถลดคุณลักษณะที่ต้องนำมาใช้

1) การค้นหาด้วยวิธีเชิงพันธุกรรม (Genetic Algorithm: GA) นำเสนอโดย Holland [11] เป็นวิธีการผสมพันธุกรรมตามธรรมชาติ (Natural Selection) จะเห็นว่าในธรรมชาติ พันธุกรรมของสิ่งมีชีวิตทุกชนิดมีการพัฒนาโดยเลือกสิ่งที่ดีที่สุดในสายพันธุ์เพื่อสืบทอดไปยังรุ่นต่อไป ในการคำนวณ ใช้จากประสบการณ์ที่มีขั้นตอนการคำนวณมาก่อนหน้าเพื่อค้นหาคำตอบที่ดีกว่าในขั้นตอนต่อไป คำตอบที่ต้องการหาถูกกำหนดในรูป Genome (หรือ Chromosome) จากนั้น GA จะสร้างและปรับปรุงคุณภาพของประชากร

(Population) โดยผ่านกระบวนการต่าง ๆ เช่น Mutation Crossover เพื่อสืบหาคำตอบในหมู่ประชากรที่ดีที่สุด จุดเด่นคือการแก้ ปัญหาที่เป็น Discrete Continuous หรือผสมได้

2) การค้นหาด้วยวิธีละโมบ (Greedy Search) เป็น การค้นหาแบบดีที่สุดก่อน หลักการของการค้นหาแบบ นี้คือเลือกเส้นทางที่ดีที่สุดก่อนเพื่อให้เข้าใกล้เป้าหมาย พิจารณาจากเส้นทางที่มองเห็นเท่านั้น โดยเลือกโหนด 1 โหนดเป็นสถานะปัจจุบันและสร้าง heuristic ($h(n)$) จากจุด n ไปยัง goal ที่ใกล้ที่สุด โดย $h(n)$ = ค่าประมาณต้นทุนจากจุด n ไป goal โดยเป็น ระยะทางที่สั้นที่สุด

การกำหนดทิศทางสำหรับการค้นหา 2 แบบดังนี้

- 1) Forward Chaining จุดเริ่มต้น + การใช้กฎ > เป้าหมาย
- 2) Backward Chaining เป้าหมาย + การใช้กฎ > จุดเริ่มต้น

2.3 การจำแนกข้อความ

การจำแนกข้อความเรียนรู้แบบมีผลเฉลย (Supervised Learning) มีขั้นตอนในการจำแนก 2 ขั้นตอนคือ การเรียนรู้เพื่อสร้างเอกสารต้นแบบและการแยกหมวดหมู่ของเอกสารที่สนใจ โดยการตรวจหา ความคล้ายกับกลุ่มเอกสารต้นแบบประกอบด้วยเทคนิค ดังต่อไปนี้

2.3.1 เบย์ (Bayes)

คือวิธีการเรียนรู้ที่ใช้หลักการของความน่าจะเป็น ซึ่งมีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes's theorem) [12] เช่นกำหนดให้การเกิดของเหตุการณ์ต่างๆ ที่ใช้ใน

การจำแนกกลุ่มนั้นเป็นอิสระต่อกัน แนวคิดทฤษฎีของ เบย์ สามารถทำนายเหตุการณ์ที่พิจารณาได้จากการเกิด ของเหตุการณ์ต่างๆ ได้ ดังสมการที่ 6

$$\Pr(class = c|X) = \frac{\Pr(X|class = c) \Pr(class = c)}{\Pr(X)} \quad (6)$$

ข้อสังเกต $\Pr(X)$ ไม่เปลี่ยนสำหรับค่าคลาส c ที่เปลี่ยนไป

$\Pr(class = c) \approx$ ความถี่สัมพัทธ์ของตัวอย่างในคลาส c

$\Pr(class = c|X)$ สูงสุดก็ต่อเมื่อ

$\Pr(X|class = c) = \Pr(class = c)$ สูงสุด

นาอ็ฟเบย์คือการใช้วิธีการของเบย์พร้อมสมมติฐาน ของการเป็นอิสระต่อกันของตัวแปรอิสระทุกตัว ดัง สมการที่ 7

$$\Pr(x_1, \dots, x_k | class = c) = \Pr(x_1 | class = c) \dots \Pr(x_k | class = c) \quad (7)$$

ถ้าลักษณะประจำ i เป็น categorical: $\Pr(x_i | class = c)$ ประเมินด้วยความถี่สัมพัทธ์ของตัวอย่างที่มีค่า x_i ในคลาส c

ถ้าลักษณะประจำ i เป็น continuous: $\Pr(x_i | class = c)$ ประเมินด้วยฟังก์ชันความหนาแน่น Gaussian

2.3.2 เคนีเรสตันเบอร์

เคนีเรสตันเบอร์ คือการตัดสินใจของคลาส สำหรับแทนเงื่อนไขหรือกรณีใหม่ โดยการตรวจสอบ จำนวนบางจำนวน หรือเงื่อนไขที่เหมือนกันหรือ ใกล้เคียงกันมากที่สุด ใช้เวลาในการคำนวณสูงเพราะ การคำนวณเป็นการเพิ่มขึ้นแบบแฟกทอเรียลตามจุด ทั้งหมด ขณะที่ Decision Tree หรือโครงข่ายประสาท เทียมประมวลผลเพื่อสร้างเงื่อนไขได้เร็วกว่า เพราะเค

เนียร์สแตนด์เนเบอร์ มีการคำนวณทุกครั้งที่มีกรณีใหม่ ดังนั้นเพื่อความรวดเร็วข้อมูลทั้งหมดที่ใช้บ่อยต้องถูกเก็บไว้ในหน่วยความจำ ชื่อว่า Memory-Based Reasoning [13]

2.3.3 ซัพพอร์ตเวกเตอร์แมชชีน (SVM)

นำเสนอโดย [14] ใช้เพื่อหาระนาบการตัดสินใจในการแบ่งข้อมูลออกเป็นสองส่วน โดยใช้สมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกันมุ่งหาผลลัพธ์ที่ดีที่สุดของการเรียนรู้ (Discriminative Training) บนการเรียนรู้จากสถิติของข้อมูล ซึ่งทำงานโดยการหาค่าระยะขอบที่มากที่สุด (Maximum Margin) ของระนาบตัดสินใจ (Decision Hyper Plane) ในการแบ่งแยกกลุ่มข้อมูลที่ใช้ฝึกฝนออกจากกัน โดยใช้ฟังก์ชันแมปข้อมูลจาก Input Space ไปยัง Feature Space และสร้างฟังก์ชันวัดความคล้ายที่เรียกว่า เคอร์เนลฟังก์ชันบน Feature Space โดยมีวัตถุประสงค์เพื่อพยายามลดความผิดพลาดจากการทำนาย (Minimize Error) พร้อมกับเพิ่มระยะแยกแยะให้มากที่สุด (Maximized Margin) ซึ่งต่างจากเทคนิคโดยทั่วไปเช่น โครงข่ายประสาทเทียม (Artificial Neural Network: ANN) ที่มุ่งเพียงทำให้ความผิดพลาดจากการทำนายให้ต่ำที่สุดเพียงอย่างเดียวเหมาะสมสำหรับข้อมูลที่มีลักษณะมิติของข้อมูลมีปริมาณมาก โดยแบ่งแยกกลุ่มจากระนาบหลายมิติให้ประสิทธิภาพที่ดีกว่าวิธีการโดยทั่วไป เคอร์เนลที่พบได้บ่อยคือโพลิโนเมียล (Polynomial) เป็นการคำนวณหาเส้นแบ่งโดยใช้สมการเชิงเส้นที่มี Degree มากกว่าสองและเรเดียลเบสฟังก์ชัน (Radial basis Function) โดยมีค่า C เป็นค่าตัวแปรที่ปรับความสมดุลระหว่างการให้ความสำคัญของระยะแยกแยะสูงสุด หรือให้ความสำคัญกับค่าความผิดพลาดที่ต้องการให้ต่ำที่สุด โดยปกติค่า C

จะกำหนดให้มีค่ามากส่วนค่า Gamma มีค่าน้อย ซึ่งสมการทั้งหมดปรากฏในหนังสือ [15] ส่วนค่าเคอร์เนลแสดงดังสมการที่ (8) และ (9) ได้แก่ Polynomial kernel: (SVMP)

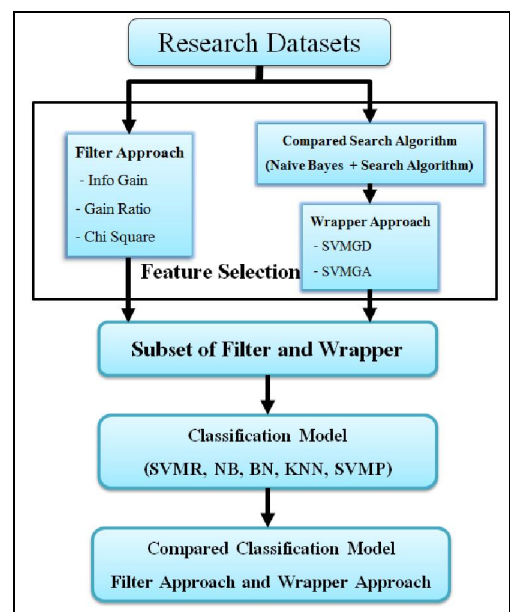
$$\gamma \times \mu' \times v + \text{coef0})^{\text{degree}} \quad (8)$$

Radial basis function kernel : (SVMR)

$$\exp(-\gamma \times |\mu - v|^2) \quad (9)$$

3. อุปกรณ์และวิธีการวิจัย

วิธีการวิจัยเริ่มต้นโดยการเตรียมข้อมูล ทำการคัดเลือกคุณลักษณะของค่าที่ดี ประกอบด้วย 2 วิธีคือการกรองและการรวบรวม ทำการจำแนกกลุ่มข้อมูลและประเมินประสิทธิภาพการจำแนกกลุ่มข้อมูลเพื่อสร้างโมเดลดังแสดงในรูปที่ 1



รูปที่ 1 โมเดลการคัดเลือกคุณลักษณะและการจำแนกข้อความ

3.1 การเตรียมข้อมูล

ข้อมูลที่ใช้ในการทดลองในครั้งนี้เป็นบทความย่อภาษาอังกฤษจากฐานข้อมูล ACM Digital Library [16] โดเมน Information System แบ่งเป็น 2 กลุ่ม คือ Database Management, Information Storage and Retrieval Information ปี 2009-2010 สกัดข้อความโดยเลือกเฉพาะคอลัมน์ที่ต้องการเช่น ชื่อผู้แต่ง ชื่อเรื่อง คำสำคัญ ฯลฯ เลือกข้อมูลที่ครบทุกคอลัมน์ และ 1 บทความมี 1 กลุ่ม เลือกคำสำคัญของเอกสารมาสร้างเป็นตัวแทนเอกสารเพื่อแทนข้อมูลทั้งหมดของเอกสาร ผลการคัดเลือกบทความจำนวน 1,009 เอกสาร ทำการสกัดข้อความโดยผ่านกระบวนการกำจัดคำหยุด การหารากศัพท์ เพื่อสร้างคำสำคัญแบบคำเดียวจำนวน = 2,354 คำ

3.2 ขั้นตอนการคัดเลือกคุณลักษณะ

การคัดเลือกคุณลักษณะใช้เพื่อลดมิติข้อมูลเนื่องจากจำนวนข้อมูลมีมากซึ่งข้อมูลที่นำมาลดมิติได้มาจากการตัดคำเดียวจำนวน 2,354 คำ ซึ่งใช้ 2 วิธี ได้แก่ การกรองและการควมรวม

3.2.1 การกรอง

การกรอง คือ การคัดเลือกคำที่มีลักษณะเฉพาะตามการคำนวณในแต่ละวิธีและได้คำที่มีอำนาจจำแนกมากที่สุด ซึ่งในการทดลองนี้ใช้สถิติ 3 วิธี ได้แก่ อินฟอร์เมชันเกน เกนเรโซ และไคสแควร์ โดยใช้ค่าการคำนวณที่ให้ผลมากกว่า 0 โดยมีรายละเอียดดังต่อไปนี้

1) อินฟอร์เมชันเกน คือการประเมินค่าเพื่อใช้ในการแบ่งข้อมูลด้วยการคำนวณค่า Gain โดยเลือกข้อมูลเฉพาะค่า IG ที่มากกว่า 0

2) เกนเรโซ คือการประเมินความน่าเชื่อถือของมิติข้อมูลโดยการวัด Gain Ratio ในแต่ละคลาส โดยเลือกข้อมูลเฉพาะค่า GR ที่มากกว่า 0

3) ไคสแควร์ คือการประเมินค่าของแอททริบิวต์โดยคำนวณค่า χ^2 เลือกข้อมูลเฉพาะค่าไคสแควร์ที่มากกว่า 0 จำนวนคำที่มีลักษณะเฉพาะของอินฟอร์เมชันเกน เกนเรโซ และไคสแควร์ มีค่า = 249 คำ

3.2.2 การควมรวม

คือ การใช้เทคนิคการทำเหมืองข้อความร่วมกับวิธีการค้นหา ในงานวิจัยนี้ทำการทดลอง 2 วิธี ดังนี้

1) เปรียบเทียบการค้นหาที่ดีที่สุดด้วยการใช้เทคนิคแบบนาอิวเบย์ ร่วมกับค้นหาแบบ Best First, Genetic Search, Greedy Stepwise, Linear Forward และ Subset Size Forward และทำการจำแนกข้อความ 5 โมเดลดังตารางที่ 1 และผลลัพธ์ดังตารางที่ 3

2) นำผลจากการเปรียบเทียบใน 1 มาใช้คือ การค้นหาด้วยวิธีเชิงพันธุกรรม (GA) และการค้นหาด้วยวิธีละโมบ (GD) ร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (SVM)

3.3 การจำแนกประเภทข้อมูล

ทำการจำแนกข้อความด้วยวิธี นาอิวเบย์ (NB) เบย์เซียนเน็ต (BN) เคเนียร์สตันเนเบอร์ (KNN) และซัพพอร์ตเวกเตอร์แมชชีน โดยใช้เคอร์เนลฟังก์ชันแบบโพลิโนเมียล (SVMP) และเรเดียลเบสิสฟังก์ชัน (SVMR) ปรับค่าพารามิเตอร์ให้เหมาะสม ซึ่งในตารางที่ 1 แสดงรายละเอียดโมเดลที่ทดลองในการจำแนกข้อความ

ตารางที่ 1 รายละเอียดโมเดลทดลองในการจำแนกข้อความ

โมเดล	รายละเอียดโมเดลที่ทดลอง	ชื่อย่อ
1.	BayesNet	BN
2.	Naive Bayes	NB
3.	K-nearest neighbor	KNN
4.	Support Vector Machine Polynomial Kernel	SVMP
5.	Support Vector Machine Radial basis Function Kernel	SVMR

งานวิจัยนี้ใช้การตรวจสอบไขว้กันหลายเท่า (K-Fold Cross Validation) เป็นวิธีการในตรวจสอบค่าความผิดพลาด ในการคาดการณ์ของโมเดล โดยพื้นฐานของวิธีการตรวจสอบไขว้กันคือการสุ่มตัวอย่าง (Resampling) ซึ่งใช้แบบ 10 Fold เพื่อใช้ในการจำแนกข้อความ

3.4 การประเมินประสิทธิภาพ

การประเมินประสิทธิภาพใช้วิธีวัดค่าความแม่นยำ (Precision: P) ค่าความระลึก (Recall: R) และการวัดประสิทธิภาพโดยรวม (F-measure) ดังสมการที่ (10) (11) และ (12) ตามลำดับ

$$P = \frac{|Ra|}{|A|} \quad (10)$$

โดยที่ $|Ra|$ คือ จำนวนข้อมูลที่ถูกต้องที่ค้นคืนออกมาได้

ตารางที่ 2 ผลการจำแนกข้อความร่วมกับการคัดเลือกคุณลักษณะแบบการกรอง

Filter	Feature	SVMR	NB	BN	KNN	SVMP
ChiSquare	249	92.2	91.7	91.4	88.7	86.5
InfoGain	249	91.2	91.5	91.4	88.7	86.5
GainRatio	249	91.0	91.5	91.4	88.7	86.5

$|A|$ คือ จำนวนข้อมูลทั้งหมดที่ค้นคืนออกมาได้

$$R = \frac{|Ra|}{|R|} \quad (11)$$

$|R|$ คือ จำนวนข้อมูลที่ถูกต้องทั้งหมดในฐานข้อมูล

$$F - measure = \frac{2 \times (R \times P)}{R + P} \quad (12)$$

4. ผลการวิจัยและการอภิปรายผล

การประเมินผลใช้วิธีวัดค่าความถูกต้องจาก ค่าความแม่นยำ ค่าความระลึก และการวัดประสิทธิภาพโดยรวม ดังสมการที่ (10) (11) และ (12) ตามลำดับสามารถอธิบายได้ 4 ขั้นตอนคือ 1) ทำการคัดเลือกคุณลักษณะแบบการกรอง 2) เปรียบเทียบการค้นหาค่าที่ 3) คัดเลือกคุณลักษณะแบบการรวบรวมโดยใช้การค้นหาค่าที่ดีจาก 2 เพื่อใช้ร่วมกับเทคนิคซอฟต์แวร์เดเตอร์แมชชีน 4) ทำการเปรียบเทียบผลลัพธ์ที่ได้จากการคัดเลือกคุณลักษณะทั้ง 2 แบบคือแบบการกรองและการรวบรวม เพื่อนำวิธีที่ดีที่สุดไปสร้างโมเดลในการลดมิติข้อมูล ซึ่งอธิบายรายละเอียดได้ตามขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 ทำการคัดเลือกคุณลักษณะแบบการกรอง 3 วิธี คือ อินฟอร์มเมชันเกน เกนเรโซ และไคสแควร์ เพื่อคัดเลือกลักษณะที่มีลักษณะเฉพาะและนำค่าเหล่านั้นมาใช้ในการจำแนกข้อความ 5 โมเดลตามตารางที่ 1 สรุปผลดังตารางที่ 2

จากตารางที่ 2 การคัดเลือกคุณลักษณะการกรองแบบไคสแควร์ ใช้วิธีการจำแนกประเภทแบบซัพพอร์ตเวกเตอร์แมชชีน โดยใช้คอร์เนลฟังก์ชันเรเดียเบสิสฟังก์ชัน (SVMR) ให้ผลการวัดประสิทธิภาพโดยรวมสูงที่สุดคือ 92.2% นาอ็ฟเบย์ 91.7% และ เบย์เซียนเนต 91.4% ตามลำดับ ซึ่งผลลัพธ์ที่ได้มีความสอดคล้องกับงานวิจัยของ Saengsiri [3] และ Haruechaiyasak [1]

ขั้นตอนที่ 2 เปรียบเทียบวิธีการค้นหาข้อมูล โดยใช้เทคนิคนาอ็ฟเบย์ร่วมกับการค้นหา แบบ Best First, Genetic Search, Greedy Stepwise, Linear Forward และ Subset Size Forward ทำการจำแนกข้อความด้วย 5 โมเดลตามตารางที่ 1 เพื่อนำวิธีการค้นหาที่ดีไปใช้งานในขั้นตอนที่ 3 ผลลัพธ์ที่ได้ดังตารางที่ 3

ตารางที่ 3 ผลลัพธ์การใช้เทคนิคแบบนาอ็ฟเบย์ร่วมกับการค้นหาข้อมูลแบบต่าง ๆ

Search	Feature	SVMR	NB	BN	KNN	SVMP
Best First	21	90.0	89.9	90.0	90.2	90.0
Genetic Search	1,261	73.2	90.0	91.0	82.4	80.9
Greedy Stepwise	21	90.1	89.8	91.1	90.2	90.0
Linear Forward	18	90.0	87.0	89.0	89.3	89.2
Subset Size	18	90.0	86.9	89.0	89.2	89.0

จากตารางที่ 3 การค้นหาข้อมูลที่ดีที่สุดคือ การค้นหาด้วยวิธีเชิงพันธุกรรม 91.0% จำนวนคำที่ใช้ 1,261 คำ และ การค้นหาด้วยวิธีละโมบ 91.1% จำนวนคำที่ใช้ 21 คำ

ขั้นตอนที่ 3 คัดเลือกคุณลักษณะแบบการควมรวม

โดยเลือกใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนร่วมกับการค้นหาด้วยวิธีเชิงพันธุกรรม (SVMGA) และการค้นหาด้วยวิธีละโมบ (SVMGD) และทำการจำแนกข้อความด้วย 5 โมเดล ตามตารางที่ 1 ผลลัพธ์ที่ได้ดังตารางที่ 4

ตารางที่ 4 ผลลัพธ์การใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนร่วมกับการค้นหาด้วยวิธีเชิงพันธุกรรม (SVMGA) และการค้นหาด้วยวิธีละโมบ (SVMGD) และทำการจำแนกข้อความด้วย 5 โมเดล

Method	Feature	SVMR	NB	BN	KNN	SVMP
SVMGD	55	94.0	89.8	90.7	91.3	92.2
SVMGA	629	77.3	86.6	87.5	85.6	84.4

* Support Vector Machine & Greedy Stepwise (SVMGD)

* Support Vector Machine & Genetic Search (SVMGA)

จากตารางที่ 4 การคัดเลือกคุณลักษณะแบบการควบรวมด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีนร่วมกับการค้นหาด้วยวิธีละโมบ (SVMGD) ใช้วิธีการจำแนกประเภทแบบซัพพอร์ตเวกเตอร์แมชชีนโดยใช้เคอร์เนลฟังก์ชันเรเดียลเบสิสฟังก์ชัน (SVMR) ให้ผลการวัดประสิทธิภาพโดยรวมสูงที่สุดคือ 94% จำนวนคำที่นำมาพิจารณาในการสร้าง คลาสมีจำนวนค่อนข้างน้อย

ขั้นตอนที่ 4 ทำการเปรียบเทียบค่าที่ได้จากการคัดเลือกคุณลักษณะทั้ง 2 แบบคือแบบการกรองและการควบรวม

ผลลัพธ์ที่ได้จากการควบรวมมีค่าการวัดประสิทธิภาพโดยรวมมากกว่าการกรอง 1.8% ซึ่งทำให้นักวิจัยสามารถนำเทคนิคของการควบรวมซึ่งเกิดจากเทคนิคการการทำเหมืองข้อความร่วมกับการค้นหาชุดข้อมูลแบบอื่น ๆ เพื่อใช้ปรับค่าความถูกต้องของข้อมูลให้เพิ่มมากยิ่งขึ้นต่อไป

5. สรุปผลและข้อเสนอแนะ

การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะแบบการกรองและการควบรวมของการทำเหมืองข้อความเพื่อการจำแนกข้อความ สรุปได้ว่าการควบรวมให้ผลการวัดค่าประสิทธิภาพโดยรวมได้ดีกว่าแบบการกรอง 1.8% โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนร่วมกับการค้นหาด้วยวิธีละโมบ (SVMGD) ทำให้ได้คำที่มีคุณลักษณะที่ดีและนำคำเหล่านั้นมาสร้างคลาสหรือซัพคลาสของออนโทโลยีตามรูปแบบการพัฒนาระบบการค้นคืนเชิงความหมายอันก่อให้เกิดการพัฒนาฐานความรู้แบบกึ่งอัตโนมัติและช่วยให้ผู้เชี่ยวชาญทำงานน้อยลงหรือง่ายขึ้น

ดังนั้นระบบการค้นคืนเชิงความหมายสามารถประยุกต์ใช้การคัดเลือกคุณลักษณะด้วยเทคนิคการจำแนกข้อความเพื่อใช้สร้างคลาสหรือซัพคลาส ซึ่งปัจจุบันงานส่วนนี้ส่วนใหญ่ต้องให้ผู้เชี่ยวชาญเป็นผู้กำหนด ส่วนการเชื่อมโยงบริบทของคำต่าง ๆ ที่เกิดขึ้นหรือมีความหมายเดียวกันสามารถเลือกใช้จากการหาค่ากฎของความสัมพันธ์ (Association Rule)

6. เอกสารอ้างอิง

- [1] C. Haruechaiyasak, W. Jitkrittum, C.Sangkeettrakarn, and C. Damrongrat, "Implementing news article category browsing based on text categorization technique", The 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI-08) workshop on Intelligent Web Interaction (IWI 2008), 2008, pp.143-146.
- [2] K. Thongklin, S. Vanichayobon and W. Wett, "Word sense disambiguation and attribute selection using gain ratio and rbf neural network", IEEE International Conference Innovation and Vision for the Future in Computing & Communication Technologies (RIVF' 08), 2008.
- [3] P. Saengsiri, P. Meesad, S. Na Wichian and U. Herwig, "Comparison of hybrid feature selection models on gene expression data", IEEE International Conference on ICT and Knowledge Engineering, 2010, pp.13 -18.

- [4] V. Nui pian, P. Meesad and P. Boonrawd, "Improve abstract data with feature selection for classification techniques", *Advanced Materials Research* 403-408, 2012, pp. 3699-3703.
- [5] P. Thamrongrat, L. Preechaveerakul and W. Wettayaprasit, "A novel voting algorithm of multi-class svm for web page classification", *The 2nd IEEE International Conference on Computer Science and Information Technology*, 2009.
- [6] R.J. Mooney and U. Nahm, "Text mining with information extraction", *Proceedings of the 4th International MIDP Colloquium*, September, 2003.
- [7] P. Meesad, V. Nui pian and P. Boonrawd, "A Chi-Square-Test for word importance differentiation in text classification" *Proceedings of 2011 International Conference on Information and Electronics Engineering (ICIEE 2011)*, 2011, pp. 110-114.
- [8] P.N. Tan, M. Steinbach, and K. Vipin, "*Introduction to Data Mining*", Addison Wesley, 2006, pp.150-163.
- [9] J.R. Quinlan, "Induction of decision trees", *Machine Learning* 1, 1968, pp. 81-106.
- [10] W. Sukontip, "Comparison of attribute selection techniques and algorithms in classifying mistaken behaviors of vocational education students", Master Thesis, Department of Computer Science, Kasetsart University, 2008. (in Thai)
- [11] J.H. Holland, "Genetic algorithm", *Scientific American* July, 1992. Available: http://www.casos.cs.cmu.edu/education/phd/classpapers/Holland_Genetic_1992.pdf, 6 March 2012.
- [12] G.L. Bretthorst, "Bayesian spectrum analysis and parameter estimation", *Lecture Notes in Statistics*, 48, Springer-Verlag, New York, 1988.
- [13] B. Links, "A detailed introduction to K-Nearest Neighbor (KNN) algorithm", Available: <http://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/>, accessed on 2 February 2011.
- [14] V. Vapnik, "*The Nature of Statistical Learning Theory*", Springer, New York, 1995.
- [15] C.J. Burges, "A tutorial on support vector machines for pattern recognition", *Data Mining and Knowledge Discovery* 2, 1998, pp. 121-167.
- [16] Association for Computing Machinery (ACM, Copyright 2009-2010), Available: <http://portal.acm.org/portal.cfm>, accessed on 4 May 2008.