การเพิ่มประสิทธิภาพการจำแนกหมวดหมู่ข้อมูลเชิงบรรณานุกรมด้านการเกษตร

Improvement of Classification for Agriculture Bibliographic Data

พิลาพรรณ โพธิ์นรินทร์[1], สุพจน์ นิตย์สุวัฒน์[2] และชูชาติ หฤไชยะศักดิ์[3]

[1]สาขาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
1518 ถ.พิบูลสงคราม แขวงวงศ์สว่าง เขตบางซื่อ กรุงเทพมหานคร 10800 โทรศัพท์ : 0-2287-9600
E-mail: pilapan.p@rmutk.ac.th

[2]ภาควิชาคอมพิวเตอร์ศึกษา คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
1518 ถ.พิบูลสงคราม แขวงวงศ์สว่าง เขตบางซื่อ กรุงเทพมหานคร 10800 โทรศัพท์ : 0-2913-2500
E-mail: sns@kmutnb.ac.th

[3]หน่วยปฏิบัติการวิจัยวิทยาการมนุษยภาษา ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ
112 ถ.พหลโยธิน ต.คลองหนึ่ง อ.คลองหลวง จ.ปทุมธานี 12120 โทรศัพท์ : 0-2564-6900
E-mail: choochart.haruechaiyasak@nectec.or.th

บทคัดย่อ

ในการจำแนกหมวดหมู่ข้อมูลเชิงบรรณานุกรม ด้วยวิธีการจำแนกหมวดหมู่โดยอาศัยชุดคำศัพท์จากพจนานุกรมเพียงอย่างเดียว อาจให้ประสิทธิภาพในการจำแนกที่ไม่ดีพอ ดังนั้นงานวิจัยนี้จึงได้นำเสนอแบบจำลองการจำแนกหมวดหมู่ข้อมูลเชิงบรรณานุกรมด้านการเกษตร เพื่อเพิ่มประสิทธิภาพในการจำแนกหมวดหมู่ ด้วยวิธีการเพิ่มชุดคำศัพท์ร่วมกับวิธีการคัดเลือกคุณลักษณะ แบบ Information Gain (IG), Chi Squared (CHI) และ Gain Ratio (GR) จากนั้นใช้อัลกอริธึมการเรียนรู้แบบมีผู้สอน (Supervised Learning) ทำการจำแนกหมวดหมู่ข้อมูล โดยใช้อัลกอริธึมต้นไม้การตัดสินใจ (Decision Tree) นาอีฟเบย์ (Naïve Bayes) และซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ข้อมูลที่ใช้ในการทดลองจำนวน 2,580 บทความ ผลลัพธ์จากการทดลองแสดงให้เห็นว่า ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ของวิธีการที่นำเสนอให้ผลลัพธ์ที่ดีกว่าการใช้ข้อมูลจากชุดคำศัพท์เดิมเพียงอย่างเดียว 1.3%

คำสำคัญ: การจำแนกหมวดหมู่, การคัดเลือกคุณลักษณะ, ข้อมูลเชิงบรรณานุกรม

Abstract

In general, classifying bibliographic data using lexicon-based might not enough in terms of classification efficiency. In this paper, we propose the agriculture bibliographic classification model by improving lexicon set and by using feature selection techniques. The techniques of Information Gain (IG), Chi Squared (CHI) and Gain Ratio (GR) are used in order to select the distinguish properties for feature selection process. Then three algorithms Decision Tree (DT), Naïve Bayes (NB) and Support Vector Machine (SVM) are applied to classify those features. The experiments were done by using 2,580 papers from agriculture publication database. The results show that the proposed method gave better performance than using only lexicon-based in terms of precision/recall and F-measure, respectively 1.3%.

Keywords: Classification, Feature Selection, Bibliographic Data

1. คำนำ

ปัจจุบันความก้าวหน้าทางเทคโนโลยีสารสนเทศ ทำให้ปริมาณข้อมูลดิจิทัลเพิ่มมากขึ้น เช่น ข้อมูลจากหน้าเว็บทั่วไป จากเว็บบล็อก (Blog) รวมทั้งข้อมูลเชิงวิชาการมีหลากหลายสาขา เช่น สาขาแพทย์ศาสตร์ มนุษยศาสตร์ เศรษฐศาสตร์ สังคมศาสตร์ วิทยาศาสตร์และเทคโนโลยี รวมทั้งสาขาการเกษตร

ศูนย์สนเทศทางการเกษตรแห่งชาติ สำนักหอสมุดมหาวิทยาลัยเกษตรศาสตร์ เป็นแหล่งรวบรวมและให้บริการสารสนเทศด้านการเกษตรของประเทศไทย และเป็นสมาชิกของเครือข่ายความร่วมมือในระบบสารสนเทศทางการเกษตรนานาชาติ ให้ความร่วมมือในการให้บริการและแลกเปลี่ยนเพื่อให้สารสนเทศด้านการเกษตรของประเทศไทยได้รับการอ้างอิงและใช้ประโยชน์ในการศึกษาค้นคว้าวิจัยในระดับนานาชาติ และเพื่อให้บุคลากรในวงการเกษตรของประเทศไทยได้ใช้ประโยชน์จากสารสนเทศด้านการเกษตรจากทั่วโลก อย่างไรก็ตามระบบที่มีอยู่ ยังขาดความสะดวกในการค้นหาผู้ที่มีความรู้ ความเชี่ยวชาญด้านการเกษตรในสาขา หรือเรื่องที่ผู้ใช้ต้องการ เพื่อตอบคำถามหรือให้คำปรึกษาแนะนำในการแก้ปัญหาต่าง ๆ ทางด้านการเกษตร เพื่อให้การจัดการและการสืบค้นข้อมูลเหล่านี้ทำได้ง่าย ต้องอาศัยการจัดแบ่งข้อมูลเป็นกลุ่มหรือหมวดหมู่ เพื่อให้การจัดเก็บและสืบค้นข้อมูลได้อย่างรวดเร็วและมีประสิทธิภาพ รองรับการสืบค้นจากผู้ใช้งานข้อมูลอย่างถูกต้องและเหมาะสม

เทคนิคการจำแนกหมวดหมู่ข้อมูล (Data Classification) เป็นเทคนิคหนึ่งที่สำคัญในการค้นพบความรู้บนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Database: KDD) [1] ปัญหาหลักของการจำแนกหมวดหมู่ข้อมูล คือ มิติของคุณลักษณะ (Feature) ของข้อมูลมีจำนวนมาก [2, 3] การคัดเลือกคุณลักษณะ (Feature Selection) เป็นขั้นตอนสำคัญในการจำแนกหมวดหมู่ข้อมูล เพื่อลดมิติของข้อมูล ทำให้ข้อมูลตั้งต้นมีขนาดเล็กลง วิธีการคัดเลือกคุณลักษณะที่ดีจะทำให้สามารถคัดเลือกคุณลักษณะที่มีความสำคัญและสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ วิธีการเลือกคุณลักษณะเพื่อทำเหมืองข้อมูล (Data Mining) มีวัตถุประสงค์เพื่อลดคุณลักษณะ เพิ่มประสิทธิภาพการพยากรณ์ เพื่อการสังเคราะห์แบบจำลองได้อย่างรวดเร็ว และเพื่อลดความซ้ำซ้อนของรูปแบบแบบจำลอง

ดังนั้นงานวิจัยนี้มีวัตถุประสงค์เพื่อเพิ่มประสิทธิภาพแบบจำลองการจำแนกหมวดหมู่ข้อมูลเชิงบรรณานุกรมด้านการเกษตร ด้วยวิธีการเพิ่มชุดคำศัพท์ร่วมกับวิธีการคัดเลือกคุณลักษณะ แบบ Information Gain (IG), Chi Squared (CHI) และ Gain Ratio (GR) โดยใช้อัลกอริธึมการจำแนกหมวดหมู่ข้อมูล ได้แก่ ต้นไม้การตัดสินใจ (Decision Tree) นาอีฟเบย์ (Naïve Bayes) และซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) วัดประสิทธิภาพจากค่าความแม่นยำ (Precision: P) ค่าความระลึก (Recall: R) และค่าความถ่วงดุล (F-measure: F) ซึ่งผลของงานวิจัยนี้จะเป็นแนวทางสำหรับการพัฒนาระบบสืบค้นผู้เชี่ยวชาญ เพื่อให้มีประสิทธิภาพในการสืบค้นผู้เชี่ยวชาญ เพื่อตอบคำถาม หรือให้คำแนะนำในการแก้ปัญหาต่าง ๆ ทางด้านการเกษตร โดยการแบ่งเนื้อหาในบทความวิจัยเป็นส่วน ๆ ดังนี้ ส่วนที่ 2 กล่าวถึงทฤษฎีและวรรณกรรมที่เกี่ยวข้อง ส่วนที่ 3 วิธีการดำเนินการวิจัย ส่วนที่ 4 ผลการดำเนินการวิจัย ส่วนที่ 5 สรุปผลการวิจัย ส่วนที่ 6 กิตติกรรมประกาศ และรายการเอกสารอ้างอิงที่ได้ศึกษา

2. ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง

งานวิจัยนี้ผู้วิจัยได้ทำการศึกษาทฤษฎีและวรรณกรรมที่เกี่ยงข้อง ได้แก่ การทำเหมืองข้อความ การจำแนกหมวดหมู่ข้อความ และวิธีการคัดเลือกคุณลักษณะ ดังนี้

2.1 การทำเหมืองข้อความ (Text Mining)

การทำเหมืองข้อความเป็นกระบวนการวิเคราะห์ข้อมูลเพื่อแยกประเภท จำแนกรูปแบบและความสัมพันธ์ใหม่ ๆ ที่เกิดขึ้นระหว่างข้อมูล และช่วยค้นพบเนื้อหาสาระที่ซ่อนอยู่ในเอกสารต่าง ๆ เพื่อนำมาใช้ประโยชน์ ในการวิจัยนี้ใช้คำสำคัญ (Keyword) มาทำการสกัดคุณลักษณะ (Feature Extraction) เพื่อดึงคุณลักษณะของคำมาเป็นตัวแทนของบทความ

2.2 การจำแนกหมวดหมู่ข้อความ (Text Categorization)

การจำแนกหมวดหมู่ข้อความ [1, 4] เป็นกระบวนการสร้างแบบจำลองจัดการข้อมูลให้อยู่ในหมวดหมู่ที่กำหนดมาให้ โดยวิธีการแบ่งกลุ่มเนื้อหาของข้อมูลที่จะใช้ที่ได้มีการกำหนดหมวดหมู่ไว้ก่อนแล้ว ซึ่งข้อมูลจะจัดให้อยู่ในกลุ่มหรือหมวดหมู่ที่คล้ายคลึงกับต้นแบบของข้อมูลนั้นมากที่สุด การสร้างแบบจำลองการจำแนกหมวดหมู่ข้อมูลซึ่งสร้างกฎการตัดสินใจจากข้อมูลที่มีอยู่โดยใช้เทคนิคดังต่อไปนี้

1) ต้นไม้การตัดสินใจ (Decision Tree: DT) เพื่อการจัดกลุ่มของ attribute input โดยใช้ Information gain ratio [5, 6]

2) นาอีฟเบย์ (Naïve Bayes: NB) คือ แบบจำลองการจำแนกกลุ่มที่ใช้หลักความน่าจะเป็นซึ่งอยู่บนพื้นฐานของ Bayes' Theorem เช่น กำหนดให้การเกิดของเหตุการณ์ต่าง ๆ ที่ใช้ในการจำแนกกลุ่มนั้นเป็นอิสระต่อกัน [7]

3) เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) คือการสร้างสมการเชิงเส้นเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกันโดย SVM พยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มให้มากที่สุด SVM ใช้ฟังก์ชันแม็ปสำหรับย้ายข้อมูลจาก Input Space ไปยัง Feature Space และสร้างฟังก์ชันวัดความคล้ายที่เรียกว่า Kernel Function บน Feature Space ข้อดีของวิธีการนี้คือ รองรับจำนวนคุณลักษณะได้มาก และมีความถูกต้องสูง ข้อเสียคือ ต้องเลือก Kernel Function ที่เหมาะสม [8]

2.3 วิธีการคัดเลือกคุณลักษณะ (Feature Selection Methods)

วิธีการคัดเลือกคุณลักษณะ [9] เป็นกระบวนการที่เลือกกลุ่มย่อยจากเซตของคุณลักษณะ (Feature set) ต้นฉบับ ซึ่งจะทำให้ได้คุณลักษณะที่เหมาะสมในการนำไปใช้ในการจำแนกหมวดหมู่ โดยที่ Feature set คือ เซตของ term หรือ word เกิดขึ้นในเอกสารทั้งหมด ซึ่งวิธีการคัดเลือกคุณลักษณะนี้จะช่วยปรับปรุงความถูกต้องในการจำแนกหมวดหมู่ของเอกสารและหลีกเลี่ยงการเกิด overfitting ได้

งานวิจัยนี้ได้เลือกวิธีการคัดเลือกคุณลักษณะที่นิยมนำมาใช้ในการจำแนกหมวดหมู่ของเอกสาร 3 วิธีดังนี้

1) Information Gain (IG) [9] ใช้การประเมินค่าของลักษณะเฉพาะโดยวัดจากค่า IG การเพิ่มคุณค่าของข้อมูลของคำสำคัญ t หาได้จากสมการที่ (1)

$$IG(t) = E(X) - \sum_{i=1}^{m} p(s_i)E(x_{Si}) \quad (1)$$

โดย

IG(t) คือการเพิ่มคุณค่าข้อมูลของคำสำคัญ t ของแอตทริบิวส์ S

$E(x)$ คือ ค่าเอนโทปี้ของแอตทริบิวต์เป้าหมาย

$S_i \in S$ คือเซ็ตย่อยของแอตทริบิวต์ S โดย $i \in \{1, 2, 3,\ldots,m\}$

m คือ เซ็ตย่อยที่เป็นไปได้ทั้งหมดของแอตทริบิวต์ S

$p(S_i)$ คือค่าความน่าจะเป็นของเซ็ตย่อย i ของแอตทริบิวต์ S

2) Chi Squared (CHI) [9] ใช้ในการทดสอบความสัมพันธ์ระหว่าง Term t และกลุ่มของหมวดหมู่ C โดยสามารถคำนวณได้ดังสมการที่ (2)

$$CHI_{avg}(t) = \sum_{i=1}^{m} p(c_i)CHI(t, c_i) \quad (2)$$

โดย

$CHI_{avg}(t)$ คือค่า Chi Squared เฉลี่ยของคำ t

$p(c_i)$ คือ ความน่าจะเป็นของหมวดหมู่ย่อยของ i

$CHI(t, c_i)$ คือค่า Chi Squared ของคำ t ที่อยู่ในหมวดหมู่ i

3) Gain Ratio (GR) [10] คือ มาตรฐานอัตราส่วนเกนใช้เพื่อเลือกคุณลักษณะที่จะใช้เป็นรากหรือโหนด แต่เนื่องจากค่ามาตรฐานเกนจะมีค่าอคติ (bias) กับข้อมูลที่ประกอบด้วยคุณสมบัติที่มีค่าเป็นไปได้จำนวนมาก จึงมีการปรับค่ามาตรฐานเกนให้ถูกต้องโดยใช้ค่าสารสนเทศของการแบ่งแยก (Split Information) ของคุณสมบัติแต่ละตัว ถ้าให้ T คือชุดของตัวอย่าง เมื่อแบ่งตัวอย่างนี้ตามคุณสมบัติของ X จะได้ ชุดของตัวอย่างย่อยในแต่ละกิ่งคือ $\{t_1, t_2,\ldots, t_n\}$ จำนวน n ชุดตามค่าที่เป็นไปได้ในคุณสมบัติ X โดยสามารถคำนวณหาค่าสารสนเทศของการแบ่งแยกเพื่อแสดงระดับการกระจายของข้อมูลได้ดังสมการที่ (3)

$$sp = -\sum_{i=1}^{n} \frac{|t_i|}{|T|} \log_2 \frac{|t_i|}{|T|} \quad (3)$$

ค่ามาตรฐานอัตราส่วนเกน = ค่ามาตรฐานเกน - ค่าสารสนเทศของการแบ่งแยก

3. วิธีการดำเนินการวิจัย

การดำเนินการวิจัยเริ่มต้นด้วยการเตรียมข้อมูลที่ใช้ในการทดลอง จากนั้นใช้คำสำคัญมาทำการสกัดคุณลักษณะของคำ เพิ่มรายการชุดคำศัพท์ในพจนานุกรม และตัดคำหยุด ในการทดลองนี้ใช้วิธีการคัดเลือกคุณลักษณะ แบบ IG, CHI และ GR อัลกอริธึมที่ใช้การจำแนกหมวดหมู่ข้อมูลได้แก่ ต้นไม้การตัดสินใจ นาอีฟเบย์ และซัพพอร์ตเวกเตอร์แมชชีน ทำการเปรียบเทียบหาประสิทธิภาพ โดยพิจารณาค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล

3.1 การเตรียมข้อมูล

งานวิจัยนี้ใช้ฐานข้อมูลบทความวิจัยด้านการเกษตรที่เผยแพร่ และตีพิมพ์ จากศูนย์สนเทศทางการเกษตรแห่งชาติ สำนักหอสมุด มหาวิทยาลัยเกษตรศาสตร์ ซึ่งเป็นฐานข้อมูลเชิงบรรณานุกรม โดยคัดเลือกเฉพาะสาขาพืช จำนวน 2,580 บทความ ภาพที่ 1 แสดงตัวอย่างบทความที่ใช้ในการทดลอง ซึ่งแต่ละบทความประกอบด้วย เมตาดาตา (Metadata) ได้แก่ ชื่อเรื่อง ชื่อเอกสาร ปีที่พิมพ์ ผู้แต่ง คำสำคัญ และบทคัดย่อ สำหรับงานวิจัยนี้ได้มีการกำหนดหมวดหมู่ให้แต่ละบทความ โดยแบ่งออกเป็น 5 หมวดหมู่ ได้แก่ การปรับปรุงพันธุ์ การใช้ปุ๋ยและการปรับปรุงดิน ศัตรูพืช โรคพืช และสภาพแวดล้อม

3.2 การวิเคราะห์ความแม่นตรงของแบบจำลอง (k-fold Cross Validation)

การตรวจสอบไขว้กัน (Cross Validation) [11] เป็นวิธีการในตรวจสอบค่าความผิดพลาด ในการคาดการณ์ของแบบจำลอง โดยพื้นฐานของวิธีการการตรวจสอบไขว้กัน คือ การสุ่มตัวอย่าง (Resampling) โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วน ๆ และนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบผลลัพธ์ โดยการตรวจสอบไขว้กันซึ่งมักถูกใช้เป็นตัวเลือกในการกำหนดแบบจำลอง

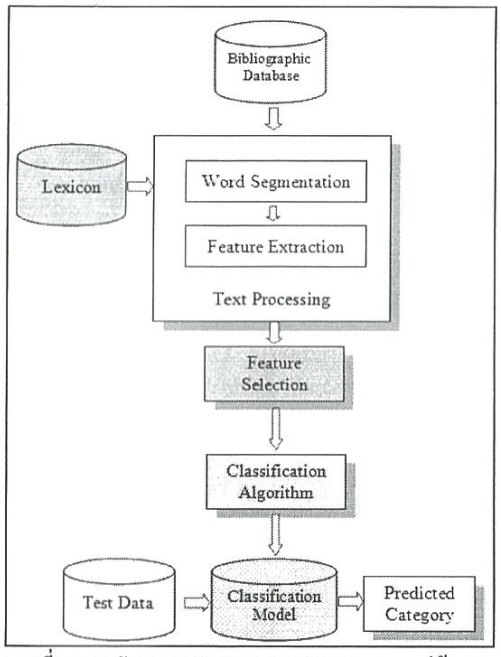
การวิเคราะห์ความแม่นตรงของแบบจำลอง เป็นการแบ่งข้อมูลออกเป็น k ชุดเท่า ๆ กันและทำการคำนวณค่าความผิดพลาด k รอบโดยแต่ละรอบการคำนวณ ข้อมูลชุดหนึ่งจากข้อมูล k ชุดจะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก k-1 ชุดจะถูกใช้เป็นข้อมูลสำหรับการเรียนรู้ แล้วนำผลความถูกต้องหรือค่าความผิดพลาดของแต่ละรอบมารวมกันและหาค่าเฉลี่ยเพื่อเป็นค่าสะท้อนประสิทธิภาพของการฝึกฝน ข้อดีของวิธีการนี้คือข้อมูลในแต่ละชุดที่ทำการแบ่งจะถูกทดสอบอย่างน้อย 1ครั้งและถูกเรียนรู้ทั้งหมด k-1 ครั้งโดยในขั้นตอนเหล่านี้สามารถกำหนดได้ว่าต้องการขนาดข้อมูลขนาดใด และต้องการทำการคำนวณเป็นจำนวนรอบเท่าใด ซึ่งเหมาะสำหรับการประมวลผลทดสอบกับข้อมูลที่มีมิติขนาดจำนวนมาก

ชื่อเรื่อง : การพัฒนาชีวภัณฑ์เชื้อปฏิปักษ์สายพันธุ์ใหม่ในการควบคุมโรคใบจุดนูนถั่วเหลือง
ชื่อเอกสาร : การประชุมทางวิชาการของมหาวิทยาลัยเกษตรศาสตร์ ครั้งที่ 43: สาขาพืช
ปีที่พิมพ์ : 2548
ผู้แต่ง : สุพจน์ กาเซ็ม Supot Kasem มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน คณะเกษตร ภาควิชาโรคพืช Kasetsart University. Bangkhen Campus, Bangkok (Thailand). Faculty of Agriculture. Department of Plant Pathology ฉวีวงศ์ สำราญอยู่ Chaweewong Samranyoo มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน คณะเกษตร ภาควิชาโรคพืช Kasetsart University. Bangkhen Campus, Bangkok (Thailand). Faculty of Agriculture. Department of Plant Pathology
คำสำคัญ : ถั่วเหลือง, โรคใบจุดนูน, เชื้อแบคทีเรียปฏิปักษ์, การควบคุมโรค, การเก็บรักษา, ความมีชีวิตรอด, ประสิทธิภาพ, การควบคุมโรคโดยชีววิธี
บทคัดย่อ : การพัฒนาสูตรสำเร็จในการเก็บเชื้อปฏิปักษ์สายพันธุ์ KPS44 KPS46 และ SW01/4 พบว่าสารพาสูตรผงโคโลไมท์ผสมรำละเอียดอัตรา 1:1 และ 1:2 รักษาความมีชีวิตของเชื้อปฏิปักษ์ได้ดีที่สุด เชื้อสายพันธุ์ KPS46 มีประชากรสูงสุดที่ 90 วัน เมื่อเก็บด้วยผงโคโลไมท์ผสมรำละเอียดอัตรา 1:1 ที่อุณหภูมิห้อง (30 องศาเซลเซียส) สายพันธุ์ KPS44 และ SW01/4 มีประชากรสูงสุดเมื่อเก็บด้วยผงโคโลไมท์ผสมรำละเอียด 1:2 ที่อุณหภูมิห้อง และ 4 องศาเซลเซียส ส่วนการเพิ่มปริมาณเชื้อปฏิปักษ์ในสารหมักชนิดต่าง ๆ พบว่าประชากรเชื้อสายพันธุ์ KPS44 สูงสุดเมื่อเลี้ยงในสารหมักสูตรปลาป่นผสมกากน้ำตาล และสูตรปลาป่นผสมกากถั่วเหลืองและกากน้ำตาลตามลำดับ สายพันธุ์ KPS46 และ SW01/4 สามารถเจริญได้ดีในสารหมักสูตรกากถั่วเหลืองผสมกากน้ำตาล ทั้งนี้สารหมักเชื้อปฏิปักษ์สูตรต่าง ๆ ที่เจือจาง 25 เปอร์เซ็นต์ สามารถลดความรุนแรงของโรคใบจุดนูนถั่วเหลืองในสภาพเรือนทดลองได้ดี สารหมักเชื้อสายพันธุ์ KPS46 สูตรปลาป่นผสมกากถั่วเหลืองและกากน้ำตาลลดความรุนแรงของโรคได้ดีที่สุด 85.7 เปอร์เซ็นต์ และไม่แตกต่างทางสถิติกับการควบคุมด้วยสาร copper hydroxide และ การใช้ suspension ของเชื้อปฏิปักษ์แต่ละสายพันธุ์

ภาพที่ 1: ตัวอย่างบทความที่ใช้ในการทดลอง

3.3 การสร้างแบบจำลองการจำแนกหมวดหมู่ข้อมูล

สร้างแบบจำลองการเรียนรู้แบบมีผู้สอน (Supervised Learning) คือ สามารถสร้างแบบจำลองการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างของข้อมูลที่ได้กำหนดได้ก่อนล่วงหน้า ที่เรียกว่า Training Set ได้อัตโนมัติ และสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้ ซึ่งได้ทำการสร้างแบบจำลอง ดังภาพที่ 2



ภาพที่ 2: แผนผังแสดงแบบจำลองการจำแนกหมวดหมู่ข้อมูลเชิงบรรณานุกรมด้านการเกษตร

จากภาพที่ 2 การทำงานเริ่มจากนำข้อมูลจากฐานข้อมูลบทความวิจัยด้านการเกษตรที่เผยแพร่และตีพิมพ์ ซึ่งเป็นข้อมูลเชิงบรรณานุกรม ตามหัวข้อ 3.1 โดยเลือกใช้คำสำคัญมาผ่านกระบวนการ ดังต่อไปนี้

1) Text Processing ประกอบด้วยขั้นตอนการทำงานดังนี้ (1) Baseline (BL) ทำการตัดคำแบบอิงชุดคำศัพท์จากพจนานุกรมโดยเลือกแบบยาวที่สุด และสกัดคุณลักษณะ เพื่อดึงคุณลักษณะของคำมาเป็นตัวแทนของบทความ และ (2) Lexicon Improvement (LI) เป็นการตัดคำแบบอิงชุดคำศัพท์จากพจนานุกรมแบบยาวที่สุด โดยเพิ่มคำศัพท์ในพจนานุกรม และทำการสกัดคุณลักษณะ ตัดคำหยุด (Stopword) ที่ไม่มีความหมายสำคัญต่อเอกสารและไม่ทำให้ใจความของเอกสารเปลี่ยน เพื่อดึงคุณลักษณะของคำมาเป็นตัวแทนของบทความ

2) Feature Selection เป็นวิธีการคัดเลือกคุณลักษณะโดยการนำคุณลักษณะของคำที่ได้จาก LI มาทำการคัดเลือกคุณลักษณะโดยใช้วิธี Information Gain (IG), Chi Squared (CHI) และ Gain Ratio (GR) เพื่อให้ได้คุณลักษณะของคำที่สำคัญและเหมาะสมในการนำไปใช้ในการจำแนกหมวดหมู่

3) Classification Algorithms การนำคุณลักษณะของคำที่ได้จาก BL, LI, IG, CHI และ GR เข้าสู่อัลกอริธึมการจำแนกหมวดหมู่ข้อมูล ได้แก่ Decision Tree, Naïve Bayes และ Support Vector Machine เพื่อสร้างแบบจำลองการจำแนกหมวดหมู่ข้อมูล ในการพยากรณ์หมวดหมู่ของข้อมูล

3.4 การทดลอง

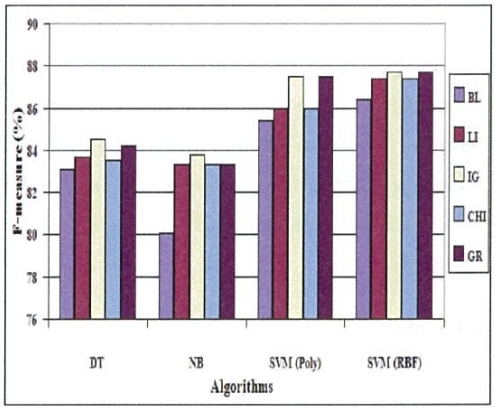
การทดลองการจำแนกหมวดหมู่ข้อมูลโดยใช้คำสำคัญมาผ่านกระบวนการ Text Processing เพื่อทำการสกัดคุณลักษณะของคำ จะได้คุณลักษณะทั้งหมดของคำ (Baseline: BL) และทำการเพิ่มคำศัพท์ในพจนานุกรม (Lexicon Improvement : LI) พร้อมทั้งเพิ่มรายการคำหยุด ที่ไม่มีความหมายสำคัญต่อเอกสารและไม่ทำให้ใจความของเอกสารเปลี่ยน เพื่อดึงคุณลักษณะของคำ จากนั้นนำคุณลักษณะทั้งหมดที่ได้จาก LI มาทำการคัดเลือกคุณลักษณะ โดยใช้วิธีการคัดเลือกคุณลักษณะ 3 วิธี คือ IG, CHI และ GR เพื่อคัดเลือกคุณลักษณะที่เหมาะสม และนำคุณลักษณะของคำทั้งหมดจาก BL และ LI รวมทั้งคุณลักษณะที่ผ่านการคัดเลือกจากทั้ง 3 วิธีเข้าสู่กระบวนการจำแนกหมวดหมู่ข้อมูล ทำการทดลองโดยใช้อัลกอริธึม ต้นไม้การตัดสินใจ นาอีฟเบย์ และซัพพอร์ตเวกเตอร์แมชชีน ผ่านทางเครื่องมือการเรียนรู้ (Machine Learning) WEKA [12] เพื่อคำนวณหาประสิทธิภาพของการจำแนกหมวดหมู่ข้อมูล

3.5 การคำนวณประสิทธิภาพ

ในการวิจัยครั้งนี้ใช้วิธีการคำนวณหาประสิทธิภาพการจำแนกหมวดหมู่ข้อมูล โดยเลือกการทดสอบประสิทธิภาพด้วยวิธี 10-fold cross validation และใช้วิธีวัดประสิทธิภาพการจำแนกหมวดหมู่ของข้อมูล จากค่าความแม่นยำ (Precision: P) ค่าความระลึก (Recall: R) และค่าความถ่วงดุล (F-measure: F) [9]

4. ผลการดำเนินการวิจัย

จากการศึกษาวิจัยพบว่า ผลลัพธ์การจำแนกหมวดหมู่ข้อมูลจากการสกัดคุณลักษณะของคำ โดยการเพิ่มชุดคำศัพท์ในพจนานุกรมพร้อมทั้งตัดคำหยุด (LI) ให้ผลลัพธ์ดีกว่าการสกัดคุณลักษณะของคำจากชุดคำศัพท์ของพจนานุกรมแบบเดิมเพียงอย่างเดียว (BL) อย่างไรก็ตามเมื่อนำเอาการสกัดคุณลักษณะแบบ LI มาทำงานร่วมกับวิธีการคัดเลือกคุณลักษณะ ได้แก่ IG, CHI และ GR ยิ่งทำให้ผลลัพธ์ของการทดลองดียิ่งขึ้น โดยเฉพาะการใช้ LI ร่วมกับ IG ให้ผลลัพธ์โดยรวมดีกว่าการใช้ LI ร่วมกับ CHI และการใช้ LI ร่วมกับ GR ทั้งนี้การทำงานร่วมกันของ LI และ IG ยังมีความเหมาะสมกับอัลกอริธึมการจำแนกหมวดหมู่ข้อมูลทั้งนาอีฟเบย์ ต้นไม้การตัดสินใจ และซัพพอร์ตเวกเตอร์แมชชีน (ที่ใช้ Kernel แบบ Polynomial และ RBF) โดยอัลกอริธึมนาอีฟเบย์ให้ค่าความถ่วงดุลเท่ากับ 83.8% ต้นไม้การตัดสินใจให้ค่าความถ่วงดุลเท่ากับ 84.5% และซัพพอร์ตเวกเตอร์แมชชีน kernel function แบบ Polynomial ให้ค่าความถ่วงดุลเท่ากับ 87.5% กับ kernel function แบบ RBF ให้ค่าความถ่วงดุลเท่ากับ 87.7% ตามลำดับ สำหรับซัพพอร์ตเวกเตอร์แมชชีน การปรับค่าพารามิเตอร์ C และ Gamma ให้เหมาะสมจะสามารถทำให้ค่าความถ่วงดุลสูงยิ่งขึ้น ดังนั้นการทดลองนี้จึงได้ทำการปรับค่าพารามิเตอร์ C ระหว่าง 1-5 และค่า Gamma ระหว่าง 0.01-0.9 ซึ่งพบว่า เมื่อค่า C มีค่าเท่ากับ 2 และค่า gamma เท่ากับ 0.4 ส่งผลให้ซัพพอร์ตเวกเตอร์แมชชีนแบบ RBF ให้ผลลัพธ์ที่ดีที่สุด แสดงดังภาพที่ 3 และตารางที่ 1



ภาพที่ 3: ผลการทดลองเปรียบเทียบค่า F-measure ระหว่าง BL, LI, IG, CHI และ GR

ตารางที่ 1: แสดงผลการจำแนกหมวดหมู่ข้อมูล

Algorithms	BL			LI			IG			CHI			GR		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
DT	83.1	83.0	**83.1**	83.8	83.6	**83.7**	84.5	84.5	**84.5**	83.7	83.4	**83.5**	84.3	84.2	**84.2**
NB	86.0	74.9	**80.1**	83.6	83.3	**83.3**	84.0	83.8	**83.8**	83.6	83.3	**83.3**	83.6	83.3	**83.3**
SVM (Poly)	83.6	87.4	**85.4**	86.1	86.0	**86.0**	87.6	87.4	**87.5**	86.1	86.0	**86.0**	87.6	87.4	**87.5**
SVM (RBF)	84.8	88.1	**86.4**	87.5	87.3	**87.4**	87.7	87.7	**87.7**	87.6	87.4	**87.4**	87.7	87.7	**87.7**

5. สรุปผลการวิจัย

การวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อนำเสนอแบบจำลองการจำแนกหมวดหมู่ เพื่อเพิ่มประสิทธิภาพในการจำแนกหมวดหมู่ข้อมูลเชิงบรรณานุกรม ด้วยวิธีการเพิ่มชุดคำศัพท์ในพจนานุกรมและตัดคำหยุด ร่วมกับวิธีการคัดเลือกคุณลักษณะ โดยใช้วิธีการคัดเลือกคุณลักษณะแบบ IG, CHI และ GR จากนั้นใช้อัลกอริธึมการเรียนรู้แบบมีผู้สอน (Supervised Learning) ทำการจำแนกหมวดหมู่ข้อมูล โดยใช้อัลกอริธึมต้นไม้การตัดสินใจ นาอีฟเบย์ และซัพพอร์ตเวกเตอร์แมชชีน ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลบทความวิจัยด้านการเกษตรที่เผยแพร่และตีพิมพ์ จากศูนย์สนเทศทางการเกษตรแห่งชาติ สำนักหอสมุด มหาวิทยาลัยเกษตรศาสตร์ ซึ่งเป็นข้อมูลเชิงบรรณานุกรม จำนวน 2,580 บทความ จากผลการทดลองพบว่า ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ของการสกัดคุณลักษณะของคำ โดยการเพิ่มชุดคำศัพท์ในพจนานุกรมพร้อมทั้งตัดคำหยุดร่วมกับวิธีการคัดเลือกคุณลักษณะให้ผลลัพธ์ที่ดีกว่าการใช้ข้อมูลจากชุดคำศัพท์เดิมเพียงอย่างเดียว 1.3%

6. กิตติกรรมประกาศ

ขอบคุณศูนย์สนเทศทางการเกษตรแห่งชาติ สำนักหอสมุด มหาวิทยาลัยเกษตรศาสตร์ที่ให้ความอนุเคราะห์ข้อมูลในการทดลอง

เอกสารอ้างอิง

[1] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2001.

[2] K. Kerdprasop, N. Kerdprasop, N. Mingmora and N. Wongprachanukul, "Wrapper and Fillter Approaches to Feature Selection in Data Mining" Proceedings of 30th Congress on Science and Technology of Thailand, Bangkok, Thailand, 2004.

[3] Yan Xu and Lin Chen, "Term-frequency based feature Selection methods for Text Categorization" The 4th International Conference on Genetic and Evolutionary Computing, pp. 280-283, 2010.

[4] F. Sebastiani, "Machine learning in automated text Categorization" ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.

[5] Youngjoong Ko and Jungyun Seo. "Using the Feature Projection Technique Based on a Normalized Voting Method for Text Classification", Information Processing & Management,vol. 40, no.2 pp. 191-208, 2004.

[6] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2005.

[7] K. Canasai and J. Chuleerat, "Thai Text Classification based on NaïveBayes", Technical Report. Department of Computer Science, Kasetsart University, 2001.

[8] Li Baoli, Lu Qin and Yu Shiwen, "An adaptive k-nearest neighbor text categorization strategy", ACM Transactions on Asian Language Information Processing (TALIP), 3 (December 2004): pp. 215-226, 2004.

[9] Yang Ying and Jan O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," Proceedings of The Fourteenth International Conference on Machine Learning (ICML'97), pp. 412-420, 1997.

[10] Cludio Ratke and Dalton Francisco de Andrade. "Using Gain Ratio Distance (GRD) to induce clustering", 5th International Conference on Intelligent Systems Design and Applications (ISDA'05), isda, pp.514-519, 2005.

[11] สมคิด แซ่หลี, "การประเมินระดับสาระเชิงหัวข้อสัมพันธ์สำหรับข้อเขียนภาษาไทยโดยเทคนิคการจัดหมวดหมู่เอกสารแบบอัตโนมัติร่วมกันออนโทโลยี", วิทยานิพนธ์ปริญญาดุษฎีบัณฑิต ภาควิชาคอมพิวเตอร์ศึกษา คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยพระจอมเกล้าพระนครเหนือ, 2550.

[12] WEKA: Waikato Environment for Knowledge Analysis, Available at: http://www.cs.waikato.ac.nz/ml/weka/.