

การจำแนกประเภทภูมิปัญญาท้องถิ่นของไทยแบบอัตโนมัติโดยวิธีการทางเหมืองข้อมูล

Automatic Thai Folk Wisdom Classification using Data Mining Approach

นพมาศ ปักเข็ม^{1*} ชนิตา จันมณี² และศิวักร อุยสุย²

Noppamas Pukkhem^{1*}, Chanida Junmanee² and Sivakorn Ouisui²

บทคัดย่อ

ในบทความนี้นำเสนอวิธีการจำแนกประเภทภูมิปัญญาท้องถิ่นแบบอัตโนมัติด้วยวิธีการทางเหมืองข้อมูล เพื่อให้สามารถจำแนกประเภทภูมิปัญญาท้องถิ่นจากข้อมูลแบบพรรณนาโวหารด้วยการประมวลผลของคอมพิวเตอร์ได้ ช่วยลดเวลาในการจับใจความรวมถึงแก้ปัญหาการจำแนกประเภทโดยมนุษย์ได้ โดยประกอบไปด้วยขั้นตอนหลัก ดังนี้ 1) การสกัดคำจากข้อความบรรยายและกำจัดคำหยุด 2) การคัดเลือกคุณลักษณะคำสำคัญ 3) การสร้างแบบจำลองการจำแนกประเภท 4) การทดสอบความแม่นยำ จากผลการทดสอบด้วยวิธี K-Folds Cross Validation ด้วยตัวอย่างข้อมูลภูมิปัญญาท้องถิ่นของไทย จำนวน 500 ข้อมูล เพื่อเปรียบเทียบความแม่นยำของการจำแนกประเภทด้วยอัลกอริทึม Decision Tree, K-Nearest Neighbor, Distance-Weight K-Nearest Neighbor และ Naïve Bayes พบว่าอัลกอริทึม Naïve Bayes ให้ค่าความแม่นยำสูงที่สุด คือ 93.5% ดังนั้นจึงนำแบบจำลองดังกล่าวมาพัฒนาส่วนติดต่อผู้ใช้งานทางหน้าเว็บด้วยภาษา PHP เชื่อมต่อกับตัวประมวลผลเหมืองข้อมูลผ่าน WEKA API เพื่ออำนวยความสะดวกในการเรียกใช้งานแบบจำลองการจำแนกประเภทข้อมูลใหม่ผ่านทางเว็บได้โดยง่าย

คำสำคัญ: การจำแนกประเภท ภูมิปัญญาท้องถิ่น เหมืองข้อมูล

Abstract

This research aims to propose the automatic Thai folk wisdom classification using data mining approaches from text description. The process consists of 4 steps, 1) word extraction and stop words elimination 2) feature selection 3) classification of model construction and 4) model evaluation. In order to compare the accuracy results of Decision Tree, K-Nearest Neighbor, Distance-Weight K-Nearest Neighbor and Naïve Bayes, K-folds cross validation technique over 500 Thai folk wisdom data were used. The comparison result of this study showed that the highest performance is Naïve Bayes with accuracy 93.5%. As this result, we selected the Naïve Bayes classifier for implementing a web-based application using WEKA API and PHP. This tool supports the ease of use when predicting the unseen data.

Keywords: Classification, Thai Folk Wisdom, Data Mining

¹ ผศ.ดร., สาขาวิชาคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยทักษิณ พัทลุง 93210

² นิสิตระดับปริญญาตรี สาขาวิชาคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยทักษิณ พัทลุง 93210

¹ Asst. Prof. Dr., Division of Computer and Information Technology, Faculty of Science, Thaksin University, Phatthalung, 93210

² Undergraduate Students, Division of Computer and Information Technology, Faculty of Science, Thaksin University, Phatthalung, 93210

* Corresponding author: E-mail addresses: n_nop@hotmail.com

บทนำ

ภูมิปัญญาท้องถิ่น มีความสำคัญต่อสังคมและประเทศชาติ เพราะเป็นเครื่องชี้ถึงความเจริญรุ่งเรืองของสังคมนั้น ซึ่งความหมายของภูมิปัญญาท้องถิ่น คือ ความสามารถในการใช้พื้นความรู้สร้างสรรค์งานเพื่อพัฒนา และดำรงชีวิตของคนในท้องถิ่น ดังนั้นถือว่าภูมิปัญญาท้องถิ่นเป็นเอกลักษณ์ของท้องถิ่นนั้น ๆ ซึ่งจะต้องมีการอนุรักษ์ และถ่ายทอดองค์ความรู้จากคนรุ่นหนึ่งไปยังรุ่นหนึ่ง

จากความสำคัญของการดำรงไว้ซึ่งภูมิปัญญาท้องถิ่น ในปัจจุบันจึงมีการเก็บข้อมูลภูมิปัญญาท้องถิ่นในรูปแบบที่หลากหลาย โดยส่วนใหญ่ มุ่งเน้นในการให้รายละเอียดข้อมูลเชิงพรรณนา ไม่ได้ระบุหมวดหมู่หรือประเภทที่ชัดเจน ซึ่งหากต่อไปมีการพัฒนาระบบการสืบค้นอาจไม่มีข้อมูลในส่วนของการจำแนกประเภท เช่น 1) เว็บไซต์ภูมิปัญญาท้องถิ่น [1] ที่มีการบอกความหมายภูมิปัญญาท้องถิ่น ลักษณะภูมิปัญญาท้องถิ่น และประเภทภูมิปัญญาท้องถิ่นที่จำแนกออกเป็นลักษณะต่างๆ 10 ลักษณะตามความรู้ของผู้จัดการข้อมูล โดยข้อมูลในเว็บไซต์จะแสดงผลเป็นไฟล์นามสกุล .pdf 2) เว็บไซต์วัฒนธรรมภูมิปัญญาท้องถิ่นภาคใต้ [2] ที่มีภูมิปัญญาในด้านต่างๆของภาคใต้ ซึ่งพบว่ามีหลากหลายในการจำแนกประเภทของภูมิปัญญาท้องถิ่นที่ต่างออกไปจากการจำแนกประเภทของกรมทรัพย์สินทางปัญญา 3) บล็อกภูมิปัญญาท้องถิ่น [3] ที่นำเสนอในรูปแบบของการเขียนบล็อกเกี่ยวกับภูมิปัญญาท้องถิ่น เรื่องสมุนไพรพื้นบ้าน และ 4) เว็บไซต์ประเพณี 4 ภาค [4] โดยการแสดงข้อมูลแบ่งตามหมวดหมู่ ประกอบไปด้วย ประเพณีท้องถิ่น วัฒนธรรมไทย การละเล่นพื้นบ้าน พิธีกรรม ความเชื่อต่าง ๆ ในงานนี้จึงได้นำแนวคิดของการจำแนกประเภท (Classification) ทางเหมืองข้อมูลมาใช้ค้นหาความรู้สำหรับการสร้างตัวแบบการจำแนกประเภทภูมิปัญญาแบบอัตโนมัติ เพื่อช่วยลดเวลา และแก้ปัญหาการจำแนกประเภทภูมิปัญญาท้องถิ่นโดยมนุษย์หรือบุคคลที่ไม่มีความเชี่ยวชาญ จากการสร้างตัวแบบการจำแนกประเภทจากข้อมูลที่สกัดความรู้จากผู้เชี่ยวชาญ เว็บไซต์และข้อมูลสารสนเทศต่าง ๆ ที่มีการบันทึกไว้ โดยจะช่วยสนับสนุนการทำงานของระบบการสืบค้นข้อมูลภูมิปัญญาท้องถิ่นมีประสิทธิภาพและเป็นระบบมากขึ้น และเป็นการช่วยส่งเสริมและอนุรักษ์ภูมิปัญญาท้องถิ่นของไทยในด้านการจัดการความรู้ โดยมีการวัดประสิทธิภาพของตัวแบบด้วยการหาค่าความแม่นยำ (Accuracy) จากข้อมูลสอนและข้อมูลทดสอบซึ่งใช้จากกรณีศึกษาภูมิปัญญาท้องถิ่นของไทย

แนวคิดและทฤษฎีที่เกี่ยวข้อง

ภูมิปัญญาท้องถิ่น

ภูมิปัญญา หรือ Wisdom หมายถึง พื้นฐานความรู้ ความสามารถ ความคิด ความเชื่อทางพฤติกรรม ความสามารถในการแก้ไขปัญหา [1] โดยใช้ประสบการณ์ที่สั่งสมไว้ในการปรับตัวและดำรงชีพในสภาพแวดล้อมทางธรรมชาติ และสิ่งแวดล้อมทางสังคมและวัฒนธรรมที่ได้มีการพัฒนาสืบสานกันมา อันเป็นผลของการใช้สติปัญญาปรับตัวให้เข้ากับสภาพต่างๆ ในพื้นที่ ที่กลุ่มชนเหล่านั้นตั้งถิ่นฐานอยู่ รวมทั้งได้มีการแลกเปลี่ยนวัฒนธรรมกับกลุ่มอื่น

จากการสำรวจข้อมูลจากเว็บไซต์ของการจำแนกภูมิปัญญาท้องถิ่นของไทย [1] มีการจำแนกประเภทภูมิปัญญาท้องถิ่นออกเป็น 9 ประเภท ดังนี้

1) ภูมิปัญญาที่เกี่ยวกับความเชื่อและศาสนา ภูมิปัญญาประเภทนี้จะมีลักษณะที่แตกต่างกันไป ในแต่ละท้องถิ่นเนื่องจากมีพื้นฐานทางความเชื่อในศาสนาที่แตกต่างกัน

2) ภูมิปัญญาท้องถิ่นที่เกี่ยวกับประเพณีและพิธีกรรม เนื่องจากประเพณีและพิธีกรรมเป็นสิ่งที่ฝังมาที่คนในท้องถิ่นสร้างขึ้นมา เป็นการเพิ่มขวัญและกำลังใจคนในสังคม จึงมีความสำคัญต่อการดำเนินชีวิตในสังคมเป็นอย่างมาก

3) ภูมิปัญญาท้องถิ่นที่เกี่ยวกับศิลปะพื้นบ้าน เป็นการสร้างสรรค์งานศิลปะต่างๆ โดยการนำ ทรัพยากรที่มีอยู่มาประยุกต์ใช้ในชีวิตประจำวัน หลังจากนั้นได้สืบทอดโดยการพัฒนาอย่างไม่ขาดสาย กลายเป็นศิลปะที่มีคุณค่าเฉพาะถิ่น

4) ภูมิปัญญาท้องถิ่นที่เกี่ยวกับอาหารและผักพื้นบ้าน นอกจากมนุษย์จะนำอาหารมาบริโภค เพื่อการอยู่รอดแล้ว มนุษย์ยังได้นำเทคนิคการถนอมอาหารและการปรุงอาหารมาใช้ เพื่อให้อาหารที่มีมากเกินไปเกินความต้องการสามารถเก็บไว้บริโภคได้เป็นเวลานาน

5) ภูมิปัญญาท้องถิ่นที่เกี่ยวกับการละเล่นพื้นบ้าน การละเล่นถือว่าเป็นการผ่อนคลายโดยเฉพาะในวัยเด็ก ซึ่งชอบความสนุกสนานเพลิดเพลิน ภูมิปัญญาท้องถิ่นของไทยส่วนใหญ่จะใช้อุปกรณ์ ในการละเล่นที่ประดิษฐ์มาจากธรรมชาติ ซึ่งแสดงให้เห็นวิถีชีวิตที่ผูกพันกับธรรมชาติและรู้จักปรับตัวให้เข้ากับสภาพแวดล้อมอย่างกลมกลืน

6) ภูมิปัญญาท้องถิ่นที่เกี่ยวกับศิลปวัฒนธรรม ประเทศไทยมีวัฒนธรรมที่หลากหลาย ซึ่งเกิดจากการสร้างสรรค์ของแต่ละภาค เราสามารถพบหลักฐานจากร่องรอยของศิลปวัฒนธรรมที่ปรากฏกระจายอยู่ทั่วไป เช่น สถาปัตยกรรม ประติมากรรม จิตรกรรม เป็นต้น ซึ่งแสดงให้เห็นถึงเทคนิค ความคิด ความเชื่อของบรรพบุรุษเป็นอย่างดี

7) ภูมิปัญญาท้องถิ่นที่เกี่ยวกับเพลงพื้นบ้าน ภูมิปัญญาประเภทนี้ส่วนมากแสดงออกถึงความ สนุกสนาน และยังเป็นคติสอนใจสำหรับคนในสังคม ซึ่งมีส่วนแตกต่างกันออกไปตามโลกทัศน์ของ คนในภาคต่างๆ

8) ภูมิปัญญาท้องถิ่นที่เกี่ยวกับสมุนไพรและตำรายาพื้นบ้าน ภูมิปัญญาประเภทนี้เกิดจากการ สังสมประสบการณ์ของคนในอดีตและถ่ายทอดให้กับคนรุ่นหลัง

9) ภูมิปัญญาท้องถิ่นที่เกี่ยวกับการประดิษฐ์กรรม เทคโนโลยีและสิ่งของเครื่องใช้ต่างๆ ที่เกิดจาก ภูมิปัญญาของคนไทยในแต่ละภาคนั้น ถือเป็นการประดิษฐ์กรรมและหัตถกรรมชั้นเยี่ยม

การจำแนกประเภท (Classification)

การทำเหมืองข้อมูล (Data Mining) คือกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น [5] ซึ่งเป็นส่วนช่วยในการจัดเก็บและตีความหมายข้อมูล จากเดิมที่มีการจัดเก็บข้อมูลอย่างง่าย มาสู่การจัดเก็บในรูปแบบฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้ มีเทคนิคที่นิยมใช้ คือ การจำแนกประเภทข้อมูล ซึ่งเป็นกระบวนการสร้างโมเดล เพื่อทำนายกลุ่มของข้อมูลใหม่ โดยข้อมูลทั้งหมดจะมีการแบ่งออกเป็น 2 กลุ่มคือกลุ่มข้อมูลสอน (Training Data) เป็นชุดข้อมูลที่มีบทบาทในการสร้างโมเดลจำแนกประเภทข้อมูลขึ้นมา และกลุ่มข้อมูลทดสอบ (Test Data) เป็นชุดข้อมูลประเมินความถูกต้องของโมเดล โดยกระบวนการสร้างตัวโมเดลจำแนกประเภทข้อมูล แบ่งออกเป็น 3 ขั้นตอน ดังนี้ 1) การสร้างโมเดลการจำแนกประเภทเป็นขั้นตอนที่อาศัยการเรียนรู้จากข้อมูลสอนที่ได้กำหนดประเภทไว้เรียบร้อยแล้ว 2) การประเมินผลโมเดล เป็นขั้นตอนตรวจสอบความถูกต้องโดยใช้ข้อมูลทดสอบ และปรับปรุงโมเดลจำแนกประเภทจนกว่าจะได้ค่าความถูกต้องในระดับที่ยอมรับได้ 3) การนำโมเดลไปใช้งาน เป็นขั้นตอนการนำโมเดลจำแนกประเภทที่สร้างขึ้นมากับข้อมูลที่ไม่เคยเห็นมาก่อนเพื่อทำนายและกำหนดกลุ่มให้กับข้อมูลนั้น

การจำแนกประเภทข้อมูลด้วยต้นไม้การตัดสินใจ (Decision Tree)

ต้นไม้การตัดสินใจ คือ กราฟต้นไม้ที่ประกอบไปด้วย จุดยอด ด้านและใบ โดยที่จุดยอดภายในแทนลักษณะประจำ (Attribute) ที่ไม่ใช่คลาสผลลัพธ์ กิ่งแทนค่าหรือผลการทดสอบ และใบแทนค่าประเภทของข้อมูลที่ต้องการจำแนก ซึ่งต้นไม้ตัดสินใจนี้ง่ายต่อการเข้าใจและการปรับเปลี่ยนเป็นกฎการจำแนก (Classification Rules) ในงานวิจัยนี้เลือกใช้อัลกอริทึม C4.5 ซึ่งใช้ค่า Gain Ratio ในการประมวลผลโดยค่า Gain Ratio [6] มีวิธีการคำนวณเช่นเดียวกับค่า Information Gain แต่จะมีการ

แก้ปัญหาในเรื่องความเอนเอียงในการคำนวณ โดยการเพิ่มการหาค่าสารสนเทศการแบ่งแยก (Split Information) [7] ซึ่งจะมีการคำนวณที่ละเอียดมากขึ้น ดังสมการที่ (3) และ (4)

$$\text{GainRatio} (S, V) = \frac{\text{Gain} (S, V)}{\text{SplitInfo} (S, V)} \quad (3)$$

$$\text{SplitInfo} (S, V) = \sum_{i=1}^m -\frac{|S_i|}{S} \times \log_2 \frac{|S_i|}{S} \quad (4)$$

เมื่อ S คือ คุณลักษณะที่ใช้ในการตัดสินใจ และ V คือ ประเภทการทำนาย

การจำแนกประเภทข้อมูลแบบเบย์อย่างง่าย (Naïve Bayes Classification)

ตัวจำแนกประเภทแบบเบย์อย่างง่าย คือ โมเดลการจำแนกประเภทข้อมูลที่ใช้หลักความน่าจะเป็นซึ่งอยู่บนพื้นฐานของ Bayes' Theorem [8] และสมมติฐานที่ให้การเกิดของเหตุการณ์ต่าง ๆ เป็นอิสระต่อกัน ถ้า กำหนดให้ P(h) ความน่าจะเป็นที่จะเกิดเหตุการณ์ h และ P(h|D) คือความน่าจะเป็นที่จะเกิดเหตุการณ์ h เมื่อเกิดเหตุการณ์ D จากตัวแปรที่กำหนด เราสามารถทำนายเหตุการณ์ที่พิจารณาจากเหตุการณ์ต่าง ๆ ได้ดังสมการที่ (5)

$$P(h | D) = [P(D | h) * P(h)] / P(D) \quad (5)$$

จากสมการข้างต้นสามารถสร้างคำนวณความน่าจะเป็นของการจำแนกประเภทแบบเบย์อย่างง่ายได้ดังสมการที่ (6)

$$P(d | h) = P(a_1, ..., a_T | h) = \prod_t P(a_t | h) \quad (6)$$

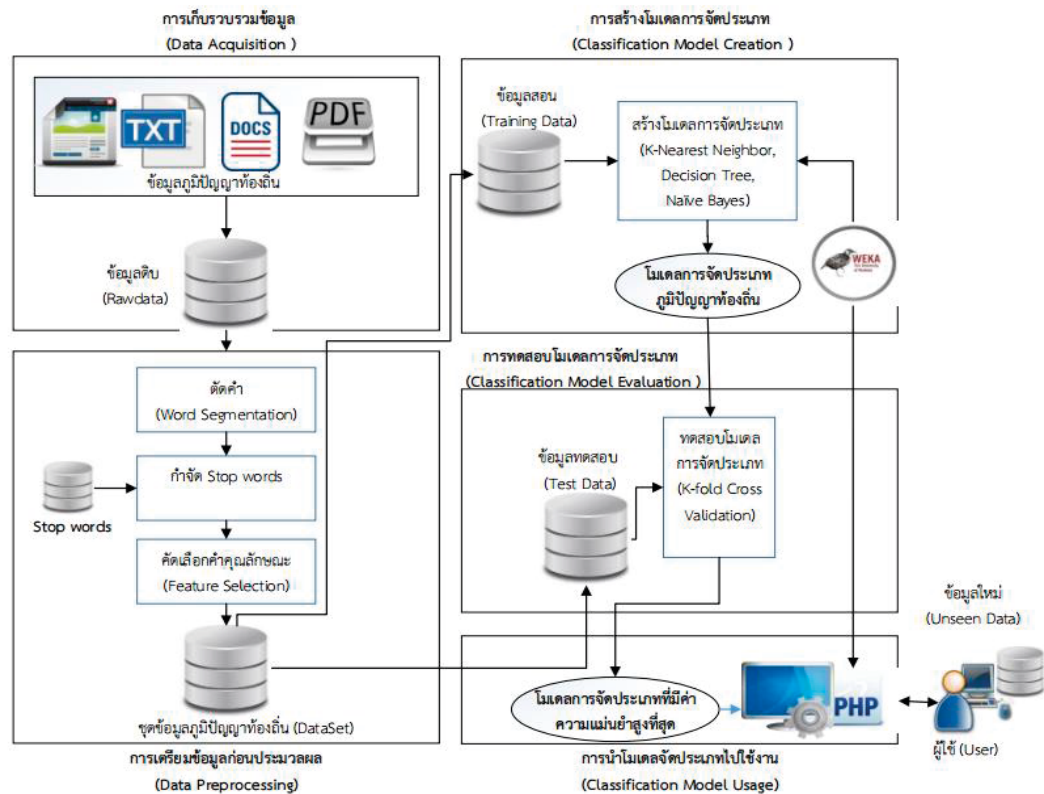
การจำแนกประเภทข้อมูลด้วยอัลกอริทึมเพื่อนบ้านใกล้สุด K ตัว (K-Nearest Neighbors-KNN)

เทคนิคของ K-NN นั้นเป็นการหาระยะห่างระหว่างแต่ละคุณลักษณะในข้อมูลและคำนวณค่าออกมา ซึ่งวิธีนี้จะเหมาะสำหรับข้อมูลแบบตัวเลข เป็นการรวมกันของผลลัพธ์ที่ได้จากการคำนวณค่าระยะห่างโดยทำการเรียงลำดับค่าระยะทางจากน้อยไปมาก หลังจากนั้นดูจากค่า “k” ว่ากำหนดเป็นเท่าไร แล้วนำลำดับที่เรียงได้มาเทียบกับคลาสข้อมูลที่เรียงแล้วนำมาตอบ [9] ระยะห่างแบบยูคลิดีเนียน (Euclidean Distance) นิยามตัวอย่าง x ด้วยเวกเตอร์คุณลักษณะ $x = \langle a_1(x), a_2(x), ..., a_n(x) \rangle$ และสามารถคำนวณระยะห่างระหว่างตัวแปรตัวอย่าง x_i และ x_j ที่แทนด้วย $d(x_i, x_j)$ ได้ดังสมการที่ (7)

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n (a_r(x_i) - a_r(x_j))^2} \quad (7)$$

ระเบียบวิธีวิจัย

ในการดำเนินงานเพื่อพัฒนากระบวนการจำแนกประเภทภูมิปัญญาท้องถิ่นแบบอัตโนมัติ มีภาพรวมการทำงานของแต่ละขั้นตอนแสดงดังภาพที่ 1



ภาพที่ 1 ภาพรวมการทำงานของระบบการจำแนกประเภทภูมิปัญญาท้องถิ่นด้วยเทคนิคทางเหมืองข้อมูล

การเก็บรวบรวมข้อมูล (Data Acquisition)

การรวบรวมข้อมูลดิบ (Raw Data) จากแหล่งสารสนเทศทางภูมิปัญญาท้องถิ่น ได้แก่ บล็อกภูมิปัญญาท้องถิ่น [3] เว็บไซต์ประเพณี 4 ภาค [4] และแหล่งข้อมูลอื่นๆ ที่เกี่ยวข้อง

การเตรียมข้อมูลก่อนการประมวลผล (Data Preprocessing)

การสร้างชุดข้อมูลภูมิปัญญาท้องถิ่น (Folk Wisdom Dataset) ประกอบไปด้วยขั้นตอนดังนี้

- 1) การสกัดคำ ในงานนี้ได้พัฒนาเครื่องมือสำหรับการตัดคำที่เชื่อมโยงฟังก์ชันเลกซ์โต (LexTo) [10] ซึ่งเป็นฟังก์ชันสำหรับการตัดคำภาษาไทยโดยใช้เทคนิคการเลือกคำที่ยาวที่สุด (Longest Matching) โดยอ้างอิงจากพจนานุกรมเล็กชิ ทรอนซ์ ด้วยการเพิ่มข้อมูลพจนานุกรมเกี่ยวกับภูมิปัญญาท้องถิ่นลงไป ตัวอย่างการตัดคำจากคำอธิบายข้อมูลภูมิปัญญาท้องถิ่น แสดงดังภาพที่ 2
- 2) การกำจัดคำที่ไม่มีผลต่อการเป็นตัวแทนของประเภทภูมิปัญญาท้องถิ่น (Stop Words Elimination)
- 3) การคัดเลือกคำคุณลักษณะ (Feature Selection)

ผลลัพธ์จากการตัดคำจะถูกนำมาหาความถี่ของการปรากฏขึ้นจากข้อมูลตัวอย่าง 500 ชุด เพื่อคัดคำที่ปรากฏบ่อยสำหรับเป็นตัวแทนประเภทของภูมิปัญญาท้องถิ่น (ประเภท A B C D E F G H และ I) หลังจากนั้นระบบจะเรียงลำดับความถี่ของคำต่างๆ โดยจะคัดเลือกคุณลักษณะคำที่ถูกใช้บ่อยที่สุด 5 ลำดับแรก พร้อมทั้งกำหนดสัญลักษณ์สำหรับการนำไปสร้างไฟล์ข้อมูลเข้าของการประมวลผลด้วยอัลกอริทึมทางเหมืองข้อมูล เช่น (A1, 20) หมายถึง “ความเชื่อ” เป็นคำที่ปรากฏขึ้น 20 ครั้งในภูมิปัญญาประเภทความเชื่อและศาสนา (A)

การจําแนกภูมิปัญญาท้องถิ่น

หนังตะลุง คือ ศิลปะการแสดงประจำท้องถิ่นอย่างหนึ่งของภาคใต้ เป็นการเล่าเรื่องราวที่ผูกร้อยเป็นนิยายดำเนินเรื่องด้วยบทหรือกรอนที่ขับร้องเป็นสำเนียงท้องถิ่น หรือที่เรียกกันว่ากรอน "ว่าบท" มีบทสนทนาแทรกเป็นระยะ และใช้การแสดงเงาบนจอผ้าเป็นสิ่งที่ดึงดูดสายตาของผู้ชม ซึ่งการว่าบท การสนทนา และการแสดงเงา นายหนังตะลุงเป็นคนแสดงเองทั้งหมด

ตัดคำ

ผลการตัดคำ

หนังตะลุง คือ ศิลปะการแสดง ประจำ ท้องถิ่น อย่างหนึ่ง ของ ภาคใต้ เป็นการ เล่าเรื่อง ราว ที่ ผูก ร้อย เป็น นิยาย ดำเนินเรื่อง ด้วย บท ร้อยกรอง ที่ ขับร้อง เป็น สำเนียง ท้องถิ่น หรือ ที่ เรียก กัน ว่ากรอน ว่าบท มี บทสนทนา แทรก เป็น ระยะ และ ใช้ การแสดง เงา บน จอผ้า เป็น สิ่ง ดึงดูด สายตา ของ ผู้ชม ซึ่ง การ ว่าบท การ สนทนา และ การแสดง เงา นี้ นาย หนังตะลุง เป็น คน แสดง เอง ทั้งหมด

ภาพที่ 2 ตัวอย่างข้อมูลเข้าและข้อมูลผลลัพธ์จากการตัดคำ

ตารางที่ 1 ผลลัพธ์การคัดเลือกคำคุณลักษณะที่ปรากฏบ่อยและการให้สัญลักษณ์เพื่อนำไปประมวลผลในอัลกอริทึมทางเหมืองข้อมูล

อันดับ	ความเชื่อและศาสนา (A)	ประเพณีและพิธีกรรม (B)	เพลงพื้นบ้าน (C)	ศิลปะพื้นบ้าน (D)	ศิลปะวัฒนธรรม (E)	การเล่นพื้นบ้าน (F)	อาหารและผักพื้นบ้าน (G)	ประเพณีธรรม (H)	สมุนไพรและตำรายาพื้นบ้าน (I)
1	ความเชื่อ (A1,20)	ประเพณี (B1,21)	เพลง (C1,15)	การแสดง (D1,9)	สันนิษฐาน (E1,9)	ละเล่น (F1,22)	อาหาร (G1,25)	ลวดลาย (H1,18)	สมุนไพร (I1,24)
2	พิธีกรรม (A2,18)	พิธีกรรม (B2,9)	ลูกคู่ (C2,11)	โนรา (D2,4)	เก่าแก่ (E2,7)	วิธีการ (F2,9)	รสชาติ (G2,18)	ความประณีต (H2,9)	ตำรา (I2,11)
3	บูชา (A3,15)	ทำบุญ (B3,8)	แม่เพลง (C3,6)	การร้อง (D3,3)	รูปแบบ (E3,6)	อุปกรณ์ (F3,9)	รับประทาน (G3,13)	सान (H3,7)	ยา (I3,6)
4	เคราะห์ (A4, 5)	ออกพรรษา (B4,5)	ท่อนเพลง (C4,5)	เครื่องดนตรี (D4,3)	จิตรกรรม (E4,6)	การเล่น (F4,9)	กิน (G4,9)	รูปทรง (H4,7)	อาการ (I4,6)
5	เทวดา (A5, 2)	ทำพิธี (B5,4)	จังหวะ (C5,5)	การรำ (D5,2)	โบราณสถาน (E5,5)	ฝ่ายตรงข้าม (F5,6)	เครื่องปรุง (G5,9)	ผลิตภัณฑ์ (H5,7)	สรรพคุณ (I5,4)

การสร้างโมเดลการจำแนกประเภท (Classification Model Creation)

หลังจากนั้นชุดข้อมูลสอนภูมิปัญญาท้องถิ่นจะถูกระบบแปลงเป็นรูปแบบไฟล์ .arff สำหรับเป็นข้อมูลเข้าในโปรแกรม WEKA [11] เพื่อนำไปทดสอบการสร้างโมเดลการจำแนกประเภทด้วยอัลกอริทึมแบบต่างๆ ได้แก่ Naïve Bayes, Decision Tree, K-Nearest Neighbor (K= 3 5 7 9)

การทดสอบโมเดลการจำแนกประเภท (Classification Model Evaluation)

นำชุดข้อมูลทดสอบเข้าทำการทดสอบความแม่นยำในการทำนายของแต่ละตัวจำแนกประเภทที่สร้างจากอัลกอริทึมประเภทต่าง ๆ โดยใช้วิธี K-Folds Cross Validation [6] คือการแบ่งข้อมูล เป็น K ส่วนเท่าๆกันแล้วทำการทดสอบ โดยเปลี่ยนข้อมูลชุดทดสอบตั้งแต่ส่วนที่ 2 เป็นชุดทดสอบและส่วนที่ 1-(K-1) เป็นชุดข้อมูลสอน ไปเรื่อยๆ จนกระทั่งถึงชุดทดสอบที่ K เป็นชุดทดสอบ และส่วนที่ 1-(K-1) เป็นชุดข้อมูลสอน โดยจะจำเอาจุดความแม่นยำที่ได้มาหาค่าเฉลี่ย

การนำโมเดลการจำแนกประเภทไปใช้งาน (Construction Model Usage)

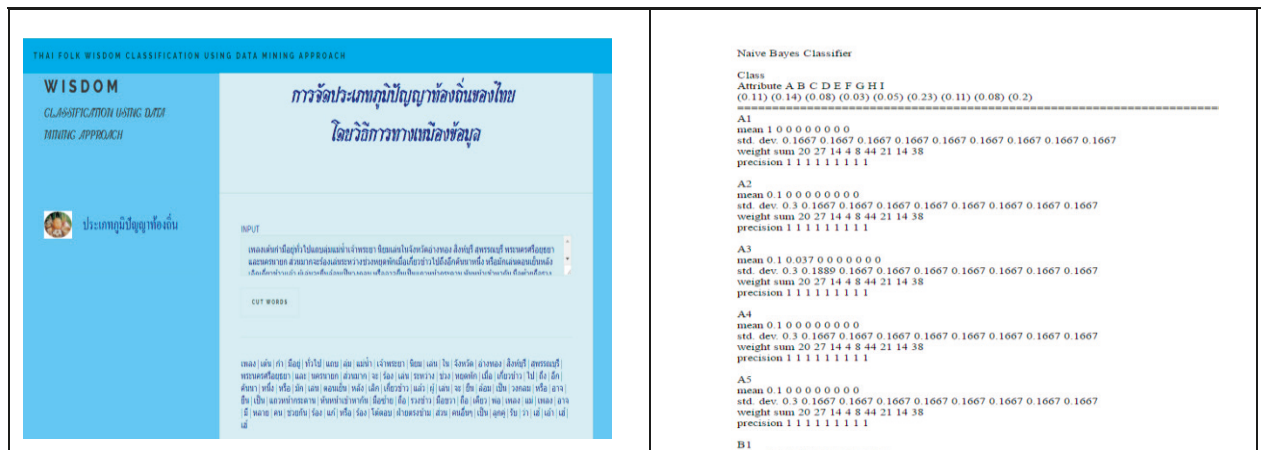
จากขั้นตอนการทดสอบโมเดล หลังจากได้โมเดลที่มีความแม่นยำสูงสุดแล้ว จะนำโมเดลดังกล่าวมาเชื่อมต่อกับส่วนติดต่อผู้ใช้ผ่านเว็บด้วย WEKA API ร่วมกับภาษา PHP ดังนั้นผู้ใช้จะสามารถทำนายข้อมูลใหม่ที่ต้องการทราบประเภทภูมิปัญญาท้องถิ่นผ่านทางเว็บได้โดยสะดวก

ผลการวิจัย

จากการใช้ข้อมูลบรรยายของภูมิปัญญาท้องถิ่นจำนวน 500 ชุด นำมาสกัดคุณลักษณะที่สำคัญและด้วยกระบวนการทางการประมวลผลค่า ได้ผลการคัดเลือกแสดงดังตารางที่ 2 นั้น จากนั้นนำมาสร้างโมเดลการจำแนกประเภท และประเมินความแม่นยำในการจัดประเภทโดยใช้ K-Folds Cross Validation ด้วยการกำหนด K= 3 5 7 9 และ 11 ตามลำดับ ผลจากการวิเคราะห์ในการทดสอบโมเดลเพื่อสร้างโมเดล แสดงดังตารางที่ 2

ตารางที่ 2 ตารางเปรียบเทียบค่าความถูกต้องของการจำแนกประเภทโดยใช้อัลกอริทึมประเภทต่าง ๆ

Classifiers	K-Folds Cross Validation (Folds)				
	3	5	7	9	11
Naïve Bayes	93.0%	93.5%	93.5%	93.5%	93.5%
Decision Tree (J48)	87.0%	88.0%	89.5%	89.0%	89.0%
K-Nearest Neighbor (k=3)	86.0%	86.5%	87.5%	86.5%	87.5%
K-Nearest Neighbor (k=5)	85.5%	85.5%	86.0%	85.5%	85.5%
K-Nearest Neighbor (k=7)	85.5%	85.5%	86.0%	85.5%	85.5%
K-Nearest Neighbor (k=9)	85.0%	85.5%	86.0%	85.5%	85.5%



ภาพที่ 3 ตัวอย่างหน้าจอของระบบจำแนกประเภทภูมิปัญญาท้องถิ่นไทยแบบอัตโนมัติ

จากผลการทดลองในตารางที่ 2 ซึ่งเป็นการนำข้อมูลการเรียนรู้มาทดสอบกับอัลกอริทึมทั้ง 3 อัลกอริทึม พบว่าอัลกอริทึม Naïve Bayes มีค่าความแม่นยำสูงสุด ที่ 93.5% และมี Root Mean Squared Error เท่ากับ 0.0892 ในงานวิจัยนี้จึงเลือกใช้อัลกอริทึม Naïve Bayes เป็นแบบจำลองโมเดล หลังจากนั้นนำโมเดล Naïve Bayes เชื่อมต่อผ่าน WEKA API มาพัฒนาส่วนติดต่อผู้ใช้ผ่านทาง Web Application ตัวอย่างหน้าจอแสดงดังภาพที่ 3

สรุปและอภิปรายผล

การจำแนกประเภทภูมิปัญญาของไทยแบบอัตโนมัติโดยวิธีการทางเหมืองข้อมูลนี้ นำเสนอวิธีการในการจำแนกประเภทภูมิปัญญาท้องถิ่นด้วยวิธีการทางเหมืองข้อมูลด้วยการประมวลผลด้วยคอมพิวเตอร์แทนมนุษย์ โดยมีการสร้างฐานความรู้ในรูปแบบตัวจำแนกประเภท ด้วย 3 อัลกอริทึม ได้แก่ Decision Tree, K-Nearest Neighbor และ Naïve Bayes พบว่าจากข้อมูลทดสอบ 500 ข้อมูล อัลกอริทึม Naive Bayes ให้ค่าความแม่นยำสูงที่สุดคือ 93.5% โดยการกำหนด 7-folds cross validation จะให้ค่าการทำนายสูงสุดในทุกโมเดล จึงเป็นค่าที่เหมาะสมต่อการแบ่งข้อมูลในการทดสอบ จากนั้นมีการพัฒนาการใช้งานโมเดลผ่านทาง Web Application โดยการพัฒนาด้วยภาษา PHP ร่วมกับ WEKA API ทำให้ผู้ใช้สามารถใช้งานระบบได้โดยง่าย ในอนาคตจะพัฒนาให้มีการเพิ่มข้อมูลสอนผ่านทาง Web Application ได้ และสามารถนำพัฒนาไปสู่ระบบแหล่งรวมสารสนเทศทางภูมิปัญญาท้องถิ่นที่มีประสิทธิภาพได้ในอนาคต

เอกสารอ้างอิง

- [1] กรมทรัพย์สินทางปัญญา. (2557). **ระบบฐานข้อมูลภูมิปัญญาท้องถิ่นไทย**. สืบค้นเมื่อ 31 มกราคม 2559, จาก <http://www.ipthailand.go.th/>
- [2] ภูมิปัญญาท้องถิ่น. (2555). สืบค้นเมื่อ 22 มกราคม 2559, จาก <http://www.corps2.com/km/05/5601151.pdf>.
- [3] บล็อกภูมิปัญญาท้องถิ่น. (2556). สืบค้นเมื่อ 28 มกราคม 2559, จาก http://supapan-kednom.blogspot.com/2008/02/blog-post_24.html.
- [4] เว็บไซต์ประเพณี 4 ภาค (2555). สืบค้นเมื่อ มกราคม 2559, จาก <http://www.prapayneethai.com/>.
- [5] บุญเสริม กิจศิริกุล. (2547). รายงานวิจัยฉบับสมบูรณ์ เรื่อง “อัลกอริทึมเหมืองข้อมูล”. ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.
- [6] Tan, P. N., Steinbach, M. and Kumar, V. (2006). **Introduction to Data Mining**. Addison-Wesley, USA.
- [7] Matthew, N. A. and Sajjan G. S. (2011). “Comparative Analysis of Serial Decision Tree Classification Algorithms”, **International Journal of Computer Science and Security**. 3, 230-240.
- [8] Richard, J. R. and Geatz, M. W. (2003). **Data Mining a Tutorial-Based Primer**. Pearson Education Inc.
- [9] Han, J., Kamber, M. and Pei, J. (2011). **Data Mining: Concepts and Techniques**. 3rd ed., Morgan Kaufmann Publishers.
- [10] เล็กซ์โต. (2012). **LexTo**. สืบค้นเมื่อ 7 กุมภาพันธ์ 2559, จาก <http://sansarn.com/lexto/>.
- [11] WEKA Tutorial. (2016). สืบค้นเมื่อ 7 กุมภาพันธ์ 2559, จาก <http://www.cs.waikato.ac.nz/ml/weka/documentation.html>.