การเปรียบเทียบประสิทธิภาพในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทย ระหว่างอัลกอริทึม LSI นาอีฟเบย์และนาอีฟเบย์ที่ปรับปรุงแล้ว

Comparison on the Performance of Thailand Tourism Web Clustering between LSI, Naïve Bayes and Modified Naïve Bayes Algorithms

นฤพนธ์ พนาวงศ์[1] และ จักรกฤษณ์ เสน่ห์ นมะหุต[2*]
Naruepon Panawong[1] and Chakkrit Snae Namahoot[2*]
[1]สาขาวิชาคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครสวรรค์ นครสวรรค์ 60000
[2]ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร พิษณุโลก 65000

บทคัดย่อ

ในงานวิจัยนี้นำเสนอการใช้อัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้วเพื่อปรับปรุงการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยตามออนโทโลจีท่องเที่ยวที่ได้ออกแบบไว้ เนื่องจากการใช้อัลกอริทึมนาอีฟเบย์ในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวมีผลให้ในแต่ละเว็บไซต์ถูกจัดให้อยู่ในหมวดหมู่เดียวเท่านั้น แต่เนื้อหาบางเว็บไซต์มีการนำเสนอเนื้อหาที่ประกอบด้วยหลายส่วนและหลากหลาย ประกอบไปด้วย สถานที่ท่องเที่ยว ที่พักและร้านอาหาร รวมอยู่ในเว็บไซต์เดียวกัน อีกทั้งจากงานวิจัยก่อนหน้านี้มีการใช้อัลกอริทึมนาอีฟเบย์ในการจัดหมวดหมู่พบว่ามีเว็บไซต์ประมาณ 130 เว็บไซต์ หรือคิดเป็นร้อยละ 27.40% จากจำนวนเว็บไซต์ทดสอบทั้งหมด 475 เว็บไซต์ จัดหมวดหมู่ไม่ถูกต้อง เช่น เว็บไซต์ร้านอาหาร ถูกจัดอยู่ในหมวดท่องเที่ยว เนื่องจากพบความถี่ของคำในหมวดสถานที่ท่องเที่ยวมากกว่าหมวดร้านอาหาร รวมถึงการใช้คำในการจัดหมวดหมู่เว็บไซต์ไม่ครอบคลุมคำบางคำที่สื่อความหมายไปในทางเดียวกันหรือคำที่มีความคล้ายคลึงกันมาในการจัดหมวดหมู่ทำให้การจัดหมวดหมู่ไม่ถูกต้อง ดังนั้นอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้วจึงได้ถูกนำมาใช้ในการวิเคราะห์เว็บไซต์เพื่อเพิ่มประสิทธิภาพในการจัดหมวดหมู่เว็บไซต์ รวมถึงเปรียบเทียบกับ Latent Semantic Indexing พร้อมวัดประสิทธิภาพของอัลกอริทึมด้วย F-Measure ซึ่งหลังจากปรับปรุงอัลกอริทึมนาอีฟเบย์แล้วพบว่ามีประสิทธิภาพดีที่สุด โดยมีค่าความแม่นยำเท่ากับ 100% ค่าความระลึกเท่ากับ 94.19% และค่า F-Measure เท่ากับ 96.58%

คำสำคัญ : จัดหมวดหมู่, การวิเคราะห์ความหมายแฝง, นาอีฟเบย์, ออนโทโลจี, วิเคราะห์เว็บ

---

* E-mail address: jnaruepon.p@gmail.com, chakkrits@nu.ac.th

Abstract

This paper presents the modified Naïve Bayes algorithm which is added to tourism ontology in order to classify tourism website in Thailand. The traditional Naïve Bayes algorithm performs better in the individual category but makes it worst for various kinds of information. In fact, results for traditional Naïve Bayes algorithm could not categorize 130 sites (27.4%) out of 475 tested pages because those web pages can be assigned to many groups. Restaurant websites, for example, are must be in attraction group because the word "restaurant" can mostly find with "dining" word. Moreover, the words which have similar meaning cannot be recognized as the same things, so that it causes for classifying incorrectness. Therefore, modified Naïve Bayes algorithm was utilized for web clustering and compared the efficiency with Latent Semantic Indexing. This approach was also tried with the F-Measure. Consequently, modified Naïve Bayes algorithm performed the best results with 100% for precision, 94.19% for recall, and 96.58% for F-Measure.

Keywords: Clustering, Latent Semantic Indexing, Naïve Bayes, Tourism Ontology, Performance

1. บทนำ

ปัจจุบันมีเว็บไซต์จำนวนมากที่ให้บริการค้นหาข้อมูลท่องเที่ยวหรือระบบที่ช่วยให้ข้อมูลเกี่ยวกับการท่องเที่ยว แต่การที่นักท่องเที่ยวไปเที่ยวยังสถานที่แห่งใดแห่งหนึ่งจำเป็นต้องรู้สถานที่ท่องเที่ยว ที่พัก ร้านอาหาร ร้านขายของฝากหรือของที่ระลึก ร้านขายสินค้าหนึ่งตำบลหนึ่งผลิตภัณฑ์ เทศกาลท่องเที่ยวประจำท้องถิ่นนั้น ๆ ซึ่งนักท่องเที่ยวส่วนใหญ่มักค้นหาผ่านทางเสริช์เอนจินอย่างกูเกิ้ล แต่ผลลัพธ์ที่ได้จากการค้นหาอาจสร้างความยากลำบากและสับสนให้แก่นักท่องเที่ยวได้ เช่น ข้อมูลที่ได้มีจำนวนมาก ข้อมูลที่ได้ไม่ตรงกับความต้องการทั้งหมด ข้อมูลที่ได้ไม่ก่อให้เกิดประโยชน์ในการนำไปใช้ ข้อมูลที่ได้ไม่ถูกจัดหมวดหมู่ไว้อย่างเหมาะสม ทำให้เสียเวลาในการคัดเลือกข้อมูลท่องเที่ยวเพื่อให้ได้ข้อมูลท่องเที่ยวที่ตรงกับความต้องการทั้งหมด อีกทั้งยังไม่สะดวกในการเข้าถึงข้อมูลท่องเที่ยวในเว็บไซต์เดียว เทคนิคการจัดหมวดหมู่เว็บไซต์ได้ถูกนำมาแก้ปัญหาเหล่านี้ โดยที่ Latent Semantic Indexing (LSI) ถูกนำไปใช้ในการจัดหมวดหมู่เว็บไซต์เกี่ยวกับสัตว์ [1] ธุรกิจ กีฬา ศิลปะ คอมพิวเตอร์แต่มีความซับซ้อนในการคำนวณ [2-3] และหาความคล้ายคลึงของเนื้อหาเว็บเพจของกลุ่มตัวอย่าง 4 เว็บไซต์ คือ การประชุมวิชาการระดับนานาชาติด้านวิศวกรรมซอฟต์แวร์และวิศวกรรมความรู้ครั้งที่ 14 เครือข่ายโรงเรียนนานาชาติด้านวิศวกรรมซอฟต์แวร์ ข้อมูลนักศึกษาด้านวิทยาศาสตร์การเมืองมหาวิทยาลัยซาเลร์โนและเว็บไซต์ Easy Clinic ซึ่ง LSI ถือเป็นการหาค่าความคล้ายคลึงระหว่างเอกสารที่ง่ายและสะดวกในการนำไปประยุกต์ใช้ในระบบค้นหาและเหมาะกับข้อมูลที่มีจำนวนมากแต่ต้องใช้เวลาในการคำนวณสูง [4] และอัลกอริทึมนาอีฟเบย์ถูกนำไปใช้ในการจัดหมวดหมู่เอกสารออนไลน์จำพวกข่าวออนไลน์ บล็อก อีเมล์และห้องสมุดดิจิตอล [5] การจัดหมวดหมู่เว็บไซต์สถานที่ท่องเที่ยว โดยใช้ข้อมูลท่องเที่ยวที่ถูกนำเสนอเป็นภาษาอังกฤษจากเว็บไซต์โลนลี่แพลนเน็ตเท่านั้น [6] จากนั้น นฤพันธ์ และจักรกฤษณ์ [7] ประยุกต์ใช้ในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยว

ประเทศไทย โดยใช้ข้อมูลท่องเที่ยวจากเว็บไดเร็กทอรี่ truehits เป็นชุดข้อมูลเรียนรู้จำนวน 1,048 เว็บไซต์ และใช้ผลลัพธ์ของการสืบค้นในเว็บไซต์กูเกิ้ลเป็นชุดข้อมูลทดสอบการจัดหมวดหมู่เว็บไซต์ท่องเที่ยว จำนวน 475 เว็บไซต์ รวมทั้งใช้คำและหมวดหมู่จากออนโทโลจีท่องเที่ยวจำนวน 6 หมวดหมู่ คือ สถานที่ท่องเที่ยว ที่พัก ร้านอาหาร ร้านขายของฝาก ร้านสินค้าหนึ่งตำบลหนึ่งผลิตภัณฑ์และเทศกาล จากนั้นนำแต่ละเว็บไซต์มาคำนวณด้วยอัลกอริทึมนาอีฟเบย์โดยใช้ค่าน้อยที่สุดในการจัดหมวดหมู่เว็บไซต์ดังกล่าวและตรวจสอบแต่ละเว็บไซต์มีการจัดหมวดหมู่ถูกต้องหรือไม่ จากการทดสอบพบว่ามีค่าความแม่นยำเท่ากับ 72.60% ค่าความระลึกเท่ากับ 70.99% และค่า F-Measure เท่ากับ 71.61% แต่อย่างไรก็ตามการใช้อัลกอริทึมนาอีฟเบย์มาทำการจัดหมวดหมู่นั้น พบว่ามีเว็บไซต์จำนวน 130 เว็บไซต์ โดยประมาณหรือ คิดเป็น 27.40% ไม่สามารถถูกจัดหมวดหมู่ได้ ทั้งนี้เนื้อหาบางเว็บไซต์มีการนำเสนอเนื้อหาที่ประกอบด้วยหลายส่วน หลากหลาย ประกอบไปด้วย สถานที่ท่องเที่ยว ที่พักและร้านอาหาร รวมอยู่ในเว็บไซต์เดียวกัน เช่น เว็บไซต์ร้านอาหาร ถูกจัดให้อยู่ในหมวดสถานที่ท่องเที่ยว เนื่องจากความถี่ของคำในหมวดสถานที่ท่องเที่ยวมากกว่าหมวดร้านอาหาร รวมถึงการใช้คำในการจัดหมวดหมู่เว็บไซต์ไม่ครอบคลุมคำบางคำที่สื่อความหมายไปในทางเดียวกันหรือคำที่มีความคล้ายคลึงกันมาในการจัดหมวดหมู่ทำให้การจัดหมวดหมู่ไม่ถูกต้อง อีกทั้งยังพบว่าเนื้อหาของเว็บไซต์มีเนื้อหาที่นำเสนอไม่ได้แสดงรายละเอียดในเรื่องใดเรื่องหนึ่งเท่านั้น ซึ่งทำให้เนื้อหาในแต่ละเว็บนั้นสามารถจัดให้อยู่ได้หลายหมวดหมู่ แต่ผลลัพธ์จากการใช้อัลกอริทึมนาอีฟเบย์จะให้ค่าความน่าจะเป็นเพียงค่าเดียวในการจัดหมวดหมู่เว็บไซต์ว่าควรถูกจัดอยู่ในหมวดหมู่ใดเพียงหมวดเดียวเท่านั้น โดยใช้ค่าการคำนวณที่น้อยสุด ซึ่งผู้วิจัยพบปัญหาคือ บางเว็บนำเสนอเนื้อหาที่หลากหลาย เช่น มีทั้งสถานที่ท่องเที่ยว ที่พัก และร้านอาหาร ในเว็บไซต์เดียว ทำให้เมื่อใช้อัลกอริทึมนี้จะทำให้เกิดความผิดพลาดในการจัดหมวดหมู่ได้ เป็นผลให้ค่าความถูกต้อง ค่าความระลึกและค่า F-Measure ลดลงตามไปด้วย

ดังนั้นผู้วิจัยจึงมีแนวคิดที่จะปรับปรุงอัลกอริทึมนาอีฟเบย์ด้วยการเพิ่มคำที่สื่อความหมายไปในทางเดียวกันหรือคำที่มีความคล้ายคลึงกันมาในการจัดหมวดหมู่ เช่น กิน ดื่ม รับประทาน หมายถึง ร้านอาหาร หลับ นอน พักผ่อน หมายถึง ที่พัก เป็นต้น และเพิ่มประสิทธิภาพในการแก้ปัญหาของเว็บไซต์ที่มีเนื้อหาครอบคลุมข้อมูลท่องเที่ยวได้หลายหมวดหมู่ โดยอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้วจะทำให้ลดข้อผิดพลาดในการวิเคราะห์เว็บไซต์และทำให้แต่ละเว็บไซต์สามารถถูกจัดให้อยู่ได้หลายหมวดหมู่ได้ อีกทั้งงานวิจัยนี้ยังได้นำเอาเทคนิค LSI มาประยุกต์ใช้ในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวอีกด้วย และนำมาผลลัพธ์ที่ได้มาทำการเปรียบเทียบประสิทธิภาพของการจัดหมวดหมู่ของเว็บไซต์ท่องเที่ยวกับอัลกอริทึมนาอีฟเบย์และอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้ว ซึ่งเนื้อหาของงานวิจัยได้แบ่งออกเป็นหัวข้อย่อยดังต่อไปนี้ วิจารณ์วรรณกรรมในหัวข้อที่ 2 วิธีดำเนินการวิจัยในหัวข้อที่ 3 ผลการวิจัยในหัวข้อที่ 4 และสรุปในหัวข้อที่ 5

2. วิจารณ์วรรณกรรม

การวิเคราะห์เว็บไซต์นั้น Gore and Pitale [8] ได้อธิบายหลักการวิเคราะห์เว็บไซต์ไว้ 3 วิธี คือ

1. วิเคราะห์เนื้อหาเว็บไซต์ ได้แก่ ข้อความ รูปภาพ เสียง วิดีโอ ที่ปรากฏอยู่ในหน้าเว็บเพจ

2. วิเคราะห์การใช้เว็บ ได้แก่ ข้อมูลที่ถูกจัดเก็บในเครื่องของผู้ใช้ เช่น Logs file, Cookies, user profile เป็นต้น

3. วิเคราะห์โครงสร้างเว็บ ได้แก่ ลิงค์ที่ปรากฏอยู่ภายในหน้าเว็บเพจ ลิงค์ที่มีการเชื่อมโยงไปยังเว็บอื่น

เกรียงกมล และจักรกฤษณ์ [9] ได้ใช้การวิเคราะห์เนื้อหาเว็บไซต์และการวิเคราะห์โครงสร้างเว็บมาใช้ในการพัฒนาระบบวิเคราะห์เว็บไซต์อนาจารด้วยกลุ่มคำเชิงความหมายคำอนาจารใน HTML Tags เพื่อป้องกันการเข้าถึงเว็บไซต์อนาจารจากบุคคลในองค์กรหรือกลุ่มครอบครัวที่มีเยาวชนอยู่ในบ้าน โดยใช้วิธีการคำนวณหาความถี่ของคำอนาจารภาษาไทยที่ปรากฏใน HTML Tags ในส่วนของหัวเว็บเพจ ตัวเว็บเพจและการเชื่อมโยงไปยังเว็บอื่น เฉพาะหน้าเริ่มต้นของเว็บไซต์เท่านั้น จากนั้นนำมาคำนวณหาค่าเรจติ้งด้วยเทคนิคทางสถิติเพื่อวิเคราะห์ว่าเป็นเว็บอนาจารหรือไม่ แต่ยังไม่สามารถวิเคราะห์เว็บไซต์ที่มีเนื้อหาบางส่วนหรือทั้งหมดเป็นภาษาอังกฤษได้ จึงทำให้ เกรียงกมล และจักรกฤษณ์ [10] พัฒนาระบบคัดกรองเว็บไซต์อนาจารด้วยเทคนิคการวิเคราะห์เว็บไซต์เพื่อแก้ปัญหาดังกล่าว พบว่ามีประสิทธิภาพสูงถึง 94% อย่างไรก็ตามเนื่องจากงานวิจัยทั้ง 2 นี้มีการนำส่วนการเชื่อมโยงไปยังเว็บอื่นมาวิเคราะห์ทำให้เกิดการทำงานที่ล่าช้า

Khan และคณะ [5] ใช้วิเคราะห์เนื้อหาเว็บไซต์สำหรับจัดกลุ่มเอกสารออนไลน์จำพวกข่าวออนไลน์ บล็อก อีเมล์และห้องสมุดดิจิตอล โดยใช้เทคนิคการทำเหมืองข้อความและกระบวนการภาษาธรรมชาติร่วมกันเพื่อให้ได้องค์ความรู้ที่ต้องการ รวมทั้งมีกระบวนการปรับเอกสารเพื่อลดความซับซ้อนของเอกสารและง่ายต่อการประมวลผล เช่น การตัดคำหยุด (Stop word) การแทนด้วยรากศัพท์ของคำ (Steming) ได้อธิบายเทคนิคในการจัดกลุ่มของเอกสารข้อความด้วยอัลกอริทึมดังต่อไปนี้ Rocchio's Algorithm, K-Nearest Neighbor, Decision Tree Decision, Rules Classification, Naïve Bayes Algorithm, Artificial Neural Network, Fuzzy Correlation, Genetic Algorithm, Support Vector Machine และวิธีผสม เช่น ระหว่าง Naïve Bayes กับ Support Vector Machine, Naïve Bayes กับ Self Organizing Map, Fuzzy k-NN เป็นต้นซึ่งพบว่าอัลกอริทึมนาอีฟเบย์ทำงานได้ดีในการกรองสแปม จัดกลุ่มอีเมล์ ตัวเลขและข้อมูลที่เป็นข้อความ ต้องการจำนวนข้อมูลที่ใช้เรียนรู้น้อย ง่ายต่อการพัฒนา ไม่ซับซ้อน เมื่อเปรียบเทียบกับอัลกอริทึมอื่น ๆ อัลกอริทึม SVM ก็เป็นอีกหนึ่งอัลกอริทึมที่เหมาะสมกับการนำไปใช้ในการจัดกลุ่มข้อมูลที่เป็นข้อความ แต่มีความยุ่งยากเรื่องพารามิเตอร์ แต่ Chau and Chen [11] ใช้อัลกอริทึม SVM ทดสอบการจัดหมวดหมู่เว็บไซต์ทางการแพทย์พบว่ามีประสิทธิภาพดีอยู่ในระดับร้อยละ 71.96

สำหรับการจำแนกประเภทของเอกสารอัลกอริทึมนาอีฟเบย์ถือว่าเป็นอัลกอริทึมที่ง่ายต่อการพัฒนาเมื่อเปรียบเทียบกับอัลกอริทึมอื่น [12] ซึ่งมีนักวิจัยต่าง ๆ ได้ใช้ประโยชน์ของอัลกอริทึมนี้ในการจัดหมวดหมู่เอกสารออนไลน์ จัดหมวดหมู่เว็บไซต์ ความคิดเห็นออนไลน์ เช่น Sureshkumar และคณะ [13] นำเสนอเทคนิคการจัดกลุ่มเอกสารอีเมล์ด้วยอัลกอริทึมนาอีฟเบย์โดยใช้การวิเคราะห์เนื้อหาเว็บไซต์แบ่งเป็น 10 กลุ่ม คือ Requisition, Complaint, Acknowledgement, Inquiry, Sales, Purchase, Goodwill, Feedback, Maths และ Computer ซึ่งในแต่ละกลุ่มนั้นจะมีคำที่ใช้ในการหาความถี่ของคำที่ปรากฏในแต่ละเอกสารอีเมล์ ซึ่งงานวิจัยนี้ชี้ให้เห็นถึงการจัดหมวดหมู่เอกสารด้วยอัลกอริทึมนาอีฟเบย์ใช้เวลาในการประมวลผลน้อยกว่า Feature Selection Method เช่น หมวดหมู่ซื้อ หมวดหมู่ขายและหมวดหมู่คำติชม อัลกอริทึมนาอีฟเบย์ใช้เวลาในการประมวลผลเท่ากับ 0.05, 0.06 และ 0.05 วินาที

ตามลำดับ แต่ Feature Selection Method ใช้เวลาในการประมวลผลเท่ากับ 0.28, 0.39 และ 0.32 วินาทีตามลำดับ

อัลกอริทึมนาอีฟเบย์ได้ถูกนำไปใช้ในการสร้างแบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยว [6] โดยใช้เทคนิคการเรียนรู้ของเครื่องโดยใช้ข้อมูลสถานที่ท่องเที่ยวจากเว็บไซต์โลนลี่แพลนเน็ตมาจัดกลุ่มสถานที่ท่องเที่ยวจำนวน 5 กลุ่มคือ สถานที่พัก สถานที่ท่องเที่ยวยามค่ำคืน ร้านอาหาร แหล่งจับจ่ายสินค้า สถานที่เยี่ยมชม แล้ววัดประสิทธิภาพของแบบจำลองทั้งหมด 4 เทคนิค คือ ต้นไม้ตัดสินใจ นาอีฟเบย์ ซัพพอร์ทเวกเตอร์แมชชีน พบว่านาอีฟเบย์ให้ค่าความถูกต้องสูงที่สุด แต่อย่างไรก็ตามแบบจำลองนี้ใช้ได้เฉพาะข้อมูลสถานที่ท่องเที่ยวที่นำเสนอเป็นภาษาอังกฤษเท่านั้น

งานวิจัยของนฤพันธ์และจักรกฤษณ์ [7] ใช้อัลกอริทึมนาอีฟเบย์ในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยแบ่งเป็น 6 หมวดหมู่ ได้แก่ สถานที่ท่องเที่ยว ที่พัก ร้านอาหาร ร้านขายของฝาก ร้านหนึ่งตำบลหนึ่งผลิตภัณฑ์และเทศกาล แต่การทำงานของอัลกอริทึมนาอีฟเบย์จะให้ค่าความน่าจะเป็นเพียงค่าเดียวเท่านั้น ทำให้เกิดข้อผิดพลาดในการจัดหมวดหมู่ เนื่องจากเนื้อหาในแต่ละเว็บไซต์มีเนื้อหาหลากหลายไม่เฉพาะเจาะจงเรื่องท่องเที่ยวเท่านั้น ยังมีเนื้อหาอื่นที่เป็นประโยชน์ต่อการท่องเที่ยว เช่น โรงแรม ที่พัก ร้านอาหาร เทศกาลต่าง ๆ รวมอยู่ในเว็บไซต์เดียวกัน

LSI ถือว่าเป็นเทคนิคหนึ่งในการประมวลผลภาษาธรรมชาติที่มีลักษณะเฉพาะในการหาความหมายที่แอบแฝงด้วยเหตุนี้จึงมีการนำมาใช้กับการสืบค้นสารสนเทศ เนื่องจากในยุคแรกนั้นใช้คำสำคัญในการระบุสิ่งที่ต้องการค้นหา แต่ปัญหาที่พบคือ บางเอกสารไม่มีคำที่ผู้ใช้ระบุแต่เอกสารนั้นตรงกับความต้องการจะไม่ถูกค้นพบ LSI นั้นเป็นวิธีการที่มีหลักการทำงานอยู่บนพื้นฐานของการคำนวณทางสถิติ โดยพิจารณาจากการปรากฏร่วมของคำต่าง ๆ ในการทำดัชนี [14] ซึ่งมีนักวิจัยนำไปประยุกต์ใช้ในงานต่าง ๆ ดังนี้ Thongbai และ Viriyapant [15] พัฒนาระบบถามตอบด้านการบริการนักศึกษาด้วย LSI เพื่อให้เกิดความสะดวกแก่ผู้ใช้ด้วยการรับคำขอและได้ผลลัพธ์ตรงตามความต้องการ โดยผู้เขียนใช้อัลกอริทึม Vector Space Model (VSM) และ LSI ในการหาความคล้ายคลึงของคำตอบที่มีในฐานข้อมูลและคำถามของผู้ใช้ จากนั้นนำมาเปรียบเทียบประสิทธิภาพพบว่าค่า F-Measure ของ LSI (95%) ดีกว่า VSM (77%) เนื่องจากสามารถใช้คำที่เป็นคำสำคัญหรือคำที่มีความหมายในเชิงเนื้อหาได้เป็นอย่างดีทำให้ได้ข้อมูลที่ตรงกับความต้องการของผู้ใช้มากกว่า สอดคล้องกับงานวิจัยของ Bhat และคณะ [16] ที่กล่าวว่าปัญหาเรื่องชื่อที่พบจากการรวบรวมข้อมูลจากเว็บไซต์มักมีการเขียนชื่อที่แตกต่างกันไป แต่หมายถึงชื่อเดียวกันหรือสถานที่เดียวกัน เช่น เมืองในประเทศอินเดีย “Bombay” บางครั้งก็พบว่าใช้ชื่อ “Mumbai” และ “บ้าน” กับ “เรือนไทย” หรือ “Human” กับ “User” มีความสัมพันธ์กันในเชิงความหมายไปในทิศทางเดียวกัน [17]

จากการวิจารณ์วรรณกรรมในงานวิจัยนี้ผู้วิจัยได้ทำการปรับปรุงอัลกอริทึมนาอีฟเบย์เพื่อทดสอบกับระบบท่องเที่ยวและเปรียบเทียบประสิทธิภาพในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยระหว่างอัลกอริทึมนาอีฟเบย์และ LSI

3. วิธีดำเนินการวิจัย

ในงานวิจัยนี้จะกล่าวถึงวิธีการของ LSI อัลกอริทึมนาอีฟเบย์และวิธีการปรับปรุงอัลกอริทึมนาอีฟเบย์และการนำไปทดสอบในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยและการวัดและเปรียบเทียบประสิทธิภาพทั้งสามวิธีการนี้

3.1 วิธีการของ LSI

ในงานวิจัยนี้ผู้วิจัยได้นำแนวคิด LSI [18] มาประยุกต์ใช้ในการวิเคราะห์เว็บไซต์เพื่อหาความคล้ายคลึงในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทย โดยมีการทำงานดังต่อไปนี้

1. นำข้อมูลท่องเที่ยวจากเว็บไดเร็กทอรี่ truehits จำนวน 1,048 เว็บแบ่งเป็น 6 หมวดหมู่ คือ สถานที่ท่องเที่ยว 233 เว็บไซต์ ที่พัก 200 เว็บไซต์ ร้านอาหาร 318 เว็บไซต์ ร้านขายของฝาก 54 เว็บไซต์ ร้านสินค้าหนึ่งตำบลหนึ่งผลิตภัณฑ์ 88 เว็บไซต์และเทศกาล 155 เว็บไซต์ มาตัด HTML Tags จากนั้นนำมาสร้างเมทริกซ์ของคำที่มีขนาด m x n ประกอบไปด้วยคำที่ใช้ในการค้นหา (m) และความถี่ของคำที่ปรากฏในแต่ละเอกสาร (n) ซึ่งคำเหล่านั้นมาจากออนโทโลจีท่องเที่ยวและจะต้องทำให้ครบทั้ง 6 หมวดหมู่ ดังแสดงตัวอย่างเมทริกซ์ของคำในหมวดสถานที่ท่องเที่ยวในรูปที่ 1 ในที่นี้ให้ n เท่ากับ 233 แต่หากข้อมูลท่องเที่ยวเพิ่มขึ้นค่า n จะถูกปรับอัตโนมัติและ m เท่ากับ 58 ตามจำนวนคำที่ใช้ในการค้นหาของหมวดสถานที่ท่องเที่ยว

M1 =

Term (m)	Web_1	Web_2	Web_3	Web_4	...	Web_{233}
เที่ยว (m_1)	431	79	152	179	...	222
น้ำตก (m_2)	15	19	26	17	...	24
ทะเล (m_3)	10	12	17	18	...	46
วัด (m_4)	8	11	68	43	...	11
:	:	:	:	:	:	:
:	:	:	:	:	:	:
ตลาด (m_{58})	13	7	19	24	....	8

รูปที่ 1. ตัวอย่างเมทริกซ์ของคำในหมวดสถานที่ท่องเที่ยว

2. เมื่อได้เมทริกซ์ของคำในแต่ละหมวดหมู่แล้ว เพื่อให้เกิดความยุติธรรมกับเอกสารที่มีความสั้นและยาวไม่เท่ากัน ผู้วิจัยจึงปรับค่าน้ำหนักของคำ โดยใช้วิธีวิเคราะห์เนื้อหาเว็บไซต์ที่เป็นข้อความในส่วนของหัวเว็บเพจ ตัวเว็บเพจ โดยไม่พิจารณาลิงค์ที่เกี่ยวข้องทั้งนี้เนื่องจากต้องการความรวดเร็วในการประมวลผล จากนั้นนำมาหาความถี่ของคำที่ปรากฏในแต่ละเว็บไซต์และค่าน้ำหนักของคำจากกลุ่มตัวอย่างเว็บไซต์ท่องเที่ยวจาก กองบรรณาธิการโปรวิชั่น [19] จำนวน 100 เว็บไซต์ แบ่งเป็นเว็บไซต์ท่องเที่ยวที่นำเสนอด้วยภาษาไทยจำนวน 50 เว็บไซต์และเว็บไซต์ท่องเที่ยวที่นำเสนอด้วยภาษาอังกฤษจำนวน 50 เว็บไซต์และหากมีเว็บไซต์เพิ่มขึ้นค่าน้ำหนักของคำจะถูกปรับใหม่อย่างอัตโนมัติ ดังแสดงตัวอย่างในตารางที่ 1 จากนั้นนำเมทริกซ์ของคำจากความถี่ (รูปที่ 1) ไปคูณกับค่าน้ำหนัก (สมการ 1) ในตาราง 1 จะได้ตามตัวอย่างในรูปที่ 2

$$\text{ค่าน้ำหนักของคำ} = \frac{\text{จำนวนเว็บไซต์ที่เข้าคู่กับคำ}}{\text{จำนวนเว็บไซต์ของกลุ่มตัวอย่าง}} \quad (1)$$

ตารางที่ 1. ตัวอย่างค่าน้ำหนักของคำ

Word	เที่ยว	น้ำตก	เกาะ	ทะเล	วัด	ภูเขา	ตลาด
Weight	1.00	0.70	0.92	0.86	0.92	0.78	0.70

M1 =

Term (m)	Web_1	Web_2	Web_3	Web_4	...	Web_n
เที่ยว (m_1)	431.00	79.00	152.00	179.00	...	222.00
น้ำตก (m_2)	10.05	13.30	18.20	11.90	...	16.80
ทะเล (m_3)	8.60	10.32	14.62	15.48	...	39.56
วัด (m_4)	7.36	10.12	62.56	39.56	...	10.12
:	:	:	:	:	:	:
:	:	:	:	:	:	:
ตลาด (m_{26})	9.10	4.90	13.30	16.80	....	5.60

รูปที่ 2. ตัวอย่างเมทริกซ์ของคำในหมวดสถานที่ท่องเที่ยวหลังคูณค่าน้ำหนักของคำ

3. สร้างเมทริกซ์สำหรับการทดสอบในแต่ละเว็บไซต์ที่มีขนาด m x n โดย n เท่ากับ 1 เนื่องจากทำทีละเว็บไซต์ โดยใช้ความถี่ของคำที่ปรากฏในเว็บไซต์ ดังแสดงตัวอย่างในรูปที่ 3 ซึ่งคำว่า “เที่ยว” ปรากฏในเว็บไซต์เป็นจำนวน 140 คำ “น้ำตก” ปรากฏในเว็บไซต์เป็นจำนวน 3 คำ “ทะเล” ปรากฏในเว็บไซต์เป็นจำนวน 5 คำ “วัด” ปรากฏในเว็บไซต์เป็นจำนวน 18 คำ และ “ตลาด” ปรากฏในเว็บไซต์เป็นจำนวน 8 คำ เป็นต้น

M2 =

Term (m)	
เที่ยว (m_1)	140
น้ำตก (m_2)	3
ทะเล (m_3)	5
วัด (m_4)	18
:	:
:	:
ตลาด (m_{58})	8

รูปที่ 3. ตัวอย่างเมทริกซ์ของคำที่ใช้ในการทดสอบ (http://travel.edtguide.com)

4. นำเมทริกซ์ที่ได้จากข้อที่ 2 มาแยกเมทริกซ์ของคำ-เอกสาร โดยใช้วิธีการ Single Value Decomposition ผลลัพธ์ที่ได้จะเป็นความสัมพันธ์ของคำ-คำ (Word-Word) คำ-เอกสาร (Word-

Document) และเอกสาร-เอกสาร (Document-Document) ตามสมการ 2 ซึ่งจะได้ค่า SVD(M1) ของหมวดสถานที่ท่องเที่ยว คือ 0.2408, 0.0250, 0.0095, -0.0141, 0.0126, -0.1058 และ -0.0467

$$SVD(M1) = U\sum V^{T} \quad (2)$$

เมื่อ U คือ เมทริกซ์เชิงตั้งฉากแทนความสัมพันธ์ของคำ-คำ

$\sum$ คือ เมทริกซ์ทแยงมุมแทนความสัมพันธ์ของคำ-เอกสาร

V^{T} คือ เมทริกซ์เชิงตั้งฉากแทนความสัมพันธ์ของเอกสาร-เอกสารและดำเนินการสับเปลี่ยนตำแหน่งตามหลักการ Transpose

5. นำเมทริกซ์ที่ได้จากข้อที่ 3 มาแยกเมทริกซ์ของคำ-เอกสาร โดยใช้วิธีการ Single Value Decomposition ผลลัพธ์ที่ได้จะเป็นความสัมพันธ์ของคำ-คำ (Word-Word) คำ-เอกสาร (Word-Document) และเอกสาร-เอกสาร (Document-Document) ตามสมการ 3 [20] ซึ่งจะได้ค่า SVD(M2) เว็บไซต์ http://travel.edtguide.com คือ 0.7276, -0.2502, 0.1717, -0.1339, 0.0786, -0.1201 และ 0.0576

$$SVD(M2) = M2^{T}U\sum^{-1} \quad (3)$$

เมื่อ $M2^{T}$ คือ เมทริกซ์ข้อมูลทดสอบแทนที่แสดงความสัมพันธ์ของคำ-เอกสารดำเนินการสับเปลี่ยนตำแหน่งตามหลักการ Transpose

U คือ เมทริกซ์เชิงตั้งฉากแทนความสัมพันธ์ของคำ-คำของข้อมูลการเรียนรู้

$\sum$ คือ เมทริกซ์ทแยงมุมแทนความสัมพันธ์ของคำ-เอกสารของข้อมูลการเรียนรู้

6. นำค่า SVD ที่ได้จากข้อที่ 4 และ 5 มาคำนวณหาค่าความคล้ายคลึง (Similarity Value) ตามสมการ 4 [21] แล้วใช้ค่าความคล้ายคลึงที่มากที่สุดในการจัดหมวดหมู่เว็บไซต์เหล่านั้น เช่น เว็บไซต์ http://travel.edtguide.com เมื่อใช้สมการ 4 แล้วจะได้ค่าความคล้ายคลึงในหมวดสถานที่ท่องเที่ยว (Att) เท่ากับ 0.84 และทำการคำนวณในแต่ละหมวดหมู่แล้วจะได้ค่าความคล้ายคลึงในหมวดที่พัก (Acc) เท่ากับ 0.63 ค่าความคล้ายคลึงในหมวดร้านอาหาร (Res) เท่ากับ 0.31 ค่าความคล้ายคลึงในหมวดร้านขายของฝาก (Sou) เท่ากับ -0.16 ค่าความคล้ายคลึงในหมวดร้านสินค้าหนึ่งตำบลหนึ่งผลิตภัณฑ์ (OTOP) เท่ากับ -0.02 และค่าความคล้ายคลึงในหมวดเทศกาล (Even) เท่ากับ 0.67 โดยงานวิจัยนี้ใช้ค่าความคล้ายคลึงที่มากที่สุดในการจัดหมวดหมู่เว็บไซต์ จะเห็นได้ว่าเว็บไซต์ http://travel.edtguide.com ถูกจัดอยู่ในหมวดสถานที่ท่องเที่ยว

$$Sim_{M1,M2} = \frac{\sum_{i=1}^{n} SVD(M1)*SVD(M2)}{\sqrt{\sum_{i=1}^{n} SVD(M1)^{2}}\sqrt{\sum_{i=1}^{n} SVD(M2)^{2}}} \quad (4)$$

เมื่อ $Sim_{M1,M2}$ คือ ค่าความคล้ายคลึงระหว่างเอกสารในแต่ละหมวดหมู่

M1 คือ เวกเตอร์ของเอกสารการเรียนรู้ในแต่ละหมวดหมู่

M2 คือ เวกเตอร์ของเอกสารทดสอบในแต่ละหมวดหมู่

7. ทำซ้ำในข้อที่ 3 จนครบทั้ง 475 เว็บไซต์

8. แสดงผลการจัดหมวดหมู่เว็บไซต์

3.2 อัลกอริทึมนาอีฟเบย์

อัลกอริทึมนาอีฟเบย์ใช้ความน่าจะเป็นในการทำนายผลโดยเป็นเทคนิคที่ใช้ในการแก้ปัญหาแบบการจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้ กระบวนการทำงานของอัลกอริทึมไม่มีความซับซ้อน [22] ซึ่งแสดงในสมการ 5 [23]

$$Cmap = \underset{c \in C}{\operatorname{argmax}} \left(P(c) \prod_{1 \le k \le n} P(t_k|c) \right) \tag{5}$$

เมื่อ Cmap คือ ค่าสูงสุดของผลคูณความน่าจะเป็นระหว่าง P(c) กับ $P(t_k \mid c)$

C คือ หมวดหมู่ทั้งหมดที่ต้องการจัด (c_1,c_2,c_3,c_4,c_5,c_6)

P(c) คือ ค่าความน่าจะเป็นในแต่ละหมวดหมู่ คำนวณได้จากสมการ 6

$P(t_k \mid c)$ คือ ค่าความน่าจะเป็นของความถี่ของคำที่ k ซึ่งปรากฏในเว็บไซต์ของหมวดหมู่ c มี 6 หมวดหมู่ คือ สถานที่ท่องเที่ยว (Att) ที่พัก (Acc) ร้านอาหาร (Res) ร้านขายของฝาก (Sou) ร้านสินค้าหนึ่งตำบลหนึ่งผลิตภัณฑ์ (OTOP) และเทศกาล (Even) สามารถคำนวณได้จากสมการ 7

n คือ จำนวนคำที่ใช้ในการจัดหมวดหมู่

$$P(c) = \frac{N_c}{N} \tag{6}$$

เมื่อ N_c คือ จำนวนเว็บไซต์เรียนรู้ที่อยู่ในหมวดหมู่ c

N คือ จำนวนเว็บไซต์เรียนรู้ทั้งหมด

$$P(t_k|c) = \frac{t_k + 1}{\sum_{k=1}^{n} t_k + 1} \tag{7}$$

แต่เมื่อคำนวณหาความน่าจะเป็นจากสมการที่ 5 แล้วจะเกิดปัญหาทศนิยมจำนวนมาก จึงได้ปรับเป็นฟังก์ชั่น log ดังสมการ 8 [23]

$$Cmap = \underset{c \in C}{\operatorname{argmax}} \left(P(c) \sum_{k=1}^{n} \log(P(t_k|c)) \right) \tag{8}$$

$$P(c)\sum_{k=1}^{n}\log(P(t_k|c))$$

โดยที่ เรียกว่า ค่าความน่าจะเป็นของผลคูณความน่าจะเป็นระหว่าง P(c) กับ $P(t_k | c)$ หรือเรียกสั้นๆ ว่า ค่าความน่าจะเป็นของนาอีฟเบย์ (P(NB))

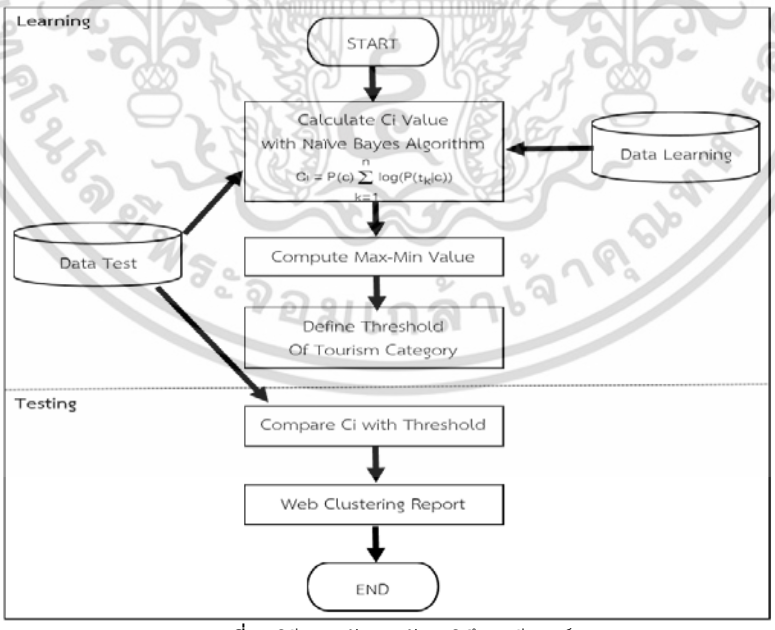
แต่เนื่องจากผลการคำนวณจากสมการ 8 ได้ค่าติดลบในทุกหมวดและจากความถี่ของคำค้นหาไปในทางกลุ่มที่มีผลการคำนวณติดลบมากที่สุด ดังนั้นผู้วิจัยจึงเลือกใช้ค่าน้อยที่สุดมาใช้ในการพิจารณาเว็บไซต์นั้นควรจัดอยู่ในหมวดหมู่ใด ดังสมการ 9

$$Cmap=\underset{c\in C}{\operatorname{argmin}}\left(P(c)\sum_{k=1}^{n}\log(P(t_k|c))\right) \quad (9)$$

ในการประยุกต์ใช้อัลกอริทึมนาอีฟเบย์ในการจัดหมวดหมู่เว็บไซต์มีขั้นตอนการทำงานดังต่อไปนี้

1. นำข้อมูลเว็บไซต์ที่ใช้สำหรับการทดสอบมาจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยด้วยอัลกอริทึมนาอีฟเบย์ (สมการ 9) จำนวน 6 หมวดหมู่ คือ สถานที่ท่องเที่ยว ที่พัก ร้านอาหาร ร้านขายของฝาก ร้านหนึ่งตำบลหนึ่งผลิตภัณฑ์และเทศกาล โดยใช้คำในออนโทโลจีท่องเที่ยวสำหรับจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทย
2. นำค่าผลคูณความน่าจะเป็นระหว่าง P(c) กับ P(tk|c) (P(NB)) ของแต่ละหมวดหมู่จากข้อที่ 1 มาเปรียบเทียบกันและเลือกค่าความน่าจะเป็นน้อยที่สุดมาใช้ในการจัดหมวดหมู่เว็บไซต์
3. ทำซ้ำในข้อที่ 1 จนกว่าจะครบทั้ง 475 เว็บไซต์
4. แสดงการจัดหมวดหมู่ของแต่ละเว็บไซต์

3.3 วิธีการปรับปรุงอัลกอริทึมนาอีฟเบย์



รูปที่ 4. วิธีการปรับปรุงอัลกอริทึมนาอีฟเบย์

ผู้วิจัยได้ทำการเพิ่มคำให้ครอบคลุมมากยิ่งขึ้นโดยใช้คำคล้ายคลึงจากพจนานุกรมสำหรับภาษาไทย และคำคล้ายคลึงจากโปรแกรมไมโครซอฟต์เวิร์ดสำหรับภาษาอังกฤษและปรับปรุงอัลกอริทึมนาอีฟเบย์ เพื่อให้ในแต่ละเว็บไซต์ถูกจัดอยู่หลายหมวดหมู่ เนื่องจากเนื้อหาในแต่ละเว็บไซต์มีเนื้อหาหลากหลายไม่เฉพาะเจาะจงเรื่องท่องเที่ยวเท่านั้น ยังมีเนื้อหาอื่นที่เป็นประโยชน์ต่อการท่องเที่ยว เช่น โรงแรม ที่พัก ร้านอาหาร เทศกาลต่าง ๆ รวมอยู่ในเว็บไซต์เดียว ซึ่งพิจารณาจากค่าความน่าจะเป็นที่อยู่ในขอบเขตของค่าเขตแดนของแต่ละหมวดหมู่ โดยมีขั้นตอนการทำงานดังต่อไปนี้ (รูปที่ 4)

1. นำข้อมูลท่องเที่ยวที่ได้จากเว็บไดเร็กทอรี่ truehits จำนวน 1,048 เว็บแบ่งเป็น 6 หมวดหมู่ คือ สถานที่ท่องเที่ยว ที่พัก ร้านอาหาร ร้านขายของฝาก ร้านหนึ่งตำบลหนึ่งผลิตภัณฑ์และเทศกาล มาตัด HTML Tags

2. นำข้อมูลในข้อที่ 1 มาคำนวณหาค่า P(NB) ของแต่ละหมวดหมู่ด้วยอัลกอริทึมนาอีฟเบย์ (สมการ 9) โดยใช้คำในออนโทโลจีท่องเที่ยวในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยว [24]

3. นำค่า P(NB) ที่ได้จากการทำงานในข้อที่ 2 ของทุกเว็บไซต์การเรียนรู้มาคำนวณหาค่า P(NB) ต่ำสุดและ P(NB) สูงสุดในแต่ละหมวดหมู่ โดยใช้สมการ 10 และ 11 จากนั้นกำหนดค่า P(NB) ต่ำสุดและ P(NB) สูงสุดในแต่ละหมวดให้เป็นค่าเขตแดน (Threshold) เพื่อใช้ในการจัดหมวดหมู่เว็บไซต์ โดยกำหนดให้ CiMax เป็นค่าสูงสุดและ CiMin เป็นค่าต่ำสุดในแต่ละหมวดหมู่ (i) ดังแสดงในตารางที่ 2

$$\text{CiMax} = \text{argmax}(\text{Cmap}) \quad 1 \le i \le 6 \tag{10}$$

$$\text{CiMin} = \text{argmin}(\text{Cmap}) \quad 1 \le i \le 6 \tag{11}$$

ตารางที่ 2. ค่าต่ำสุดและสูงสุดในแต่ละหมวดหมู่จากอัลกอริทึมนาอีฟเบย์จำนวน 1,048 เว็บไซต์

หมวดหมู่	CiMin	CiMax
สถานที่ท่องเที่ยว (Att)	-20.06	-0.08
ที่พัก (Acc)	-4.35	-0.08
ร้านอาหาร (Res)	-14.97	-0.13
ร้านขายของที่ระลึก (Sou)	-0.47	-0.02
ร้านสินค้าหนึ่งตำบลหนึ่งผลิตภัณฑ์ (OTOP)	-1.17	-0.03
เทศกาล (Even)	-4.21	-0.05

4. นำข้อมูลเว็บไซต์ที่ใช้สำหรับทดสอบมาคำนวณหาค่า P(NB)ของทุกหมวดหมู่

5. นำค่า P(NB) ที่ได้จากข้อที่ 4 ทุกหมวดหมู่มาทำการเปรียบเทียบกับค่าต่ำสุดและสูงสุดในตารางที่ 2 โดยมีเงื่อนไขดังต่อไปนี้

5.1 ถ้าค่า P(NB) ที่ได้จากการคำนวณมีค่าอยู่ในช่วงค่าต่ำสุดและค่าสูงสุดที่กำหนดไว้ในตารางที่ 2 แสดงว่าเว็บไซต์นี้สามารถจัดให้อยู่ในหมวดหมู่นั้นได้ แต่ถ้าค่า P(NB) ไม่อยู่ในช่วงที่กำหนดไว้ก็ไม่สามารถจัดให้อยู่ในหมวดหมู่นั้นได้

5.2 ทำซ้ำในข้อที่ 5.1 จนครบทั้ง 475 เว็บไซต์

6. แสดงผลการจัดหมวดหมู่เว็บไซต์

ตัวอย่างการทดสอบอัลกอริทึมนาอีฟเบย์ที่ปรับปรุง

ในงานวิจัยนี้ใช้เว็บไซต์ท่องเที่ยวจำนวน 475 เว็บไซต์สำหรับทดสอบการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทย ซึ่งผู้วิจัยขอยกตัวอย่างเว็บไซต์ travel.edtguide.com ในการแสดงวิธีทดสอบการทำงานของอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้ว เริ่มจากหาความถี่ของคำที่ปรากฏในเนื้อหาเว็บไซต์นี้ดังแสดงในตารางที่ 3

ตารางที่ 3. ตัวอย่างความถี่ของคำที่ถูกค้นพบในเว็บไซต์ travel.edtguide.com

Word	เที่ยว	น้ำตก	เกาะ	วัด	ที่พัก	ร้านอาหาร	เทศกาล
Frequency	140	3	10	15	29	12	22

จากนั้นนำค่าความถี่ไปคำนวณหาค่า P(NB) จากสมการ 9 ในแต่ละหมวดหมู่ ซึ่งหมวดสถานที่ท่องเที่ยว (Att) ได้ค่า P(NB) ของหมวดสถานที่ท่องเที่ยว (Att) เท่ากับ -4.12 หมวดที่พัก (Acc) ได้ค่า P(NB) เท่ากับ -0.53 หมวดร้านอาหาร (Res) ได้ค่า P(NB) เท่ากับ -6.00 หมวดร้านขายของฝาก (Sou) ได้ค่า P(NB) เท่ากับ 0.00 หมวดร้านสินค้าหนึ่งตำบลหนึ่งผลิตภัณฑ์ (OTOP) ได้ค่า P(NB) เท่ากับ 0.00 และหมวดเทศกาล (Even) ได้ค่า P(NB) เท่ากับ -1.61 จากนั้นนำค่า P(NB) ในแต่ละหมวดหมู่ไปเปรียบเทียบในตารางที่ 2 จะเห็นได้ว่าเว็บไซต์ travel.edtguide.com สามารถจัดให้อยู่ใน 4 หมวดหมู่ คือ สถานที่ท่องเที่ยว ที่พัก ร้านอาหารและเทศกาล ตามลำดับ

จากนั้นนำเว็บไซต์มาทดสอบให้ครบ 475 เว็บไซต์ โดยผลลัพธ์ของการคำนวณด้วยอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้วแสดงในตารางที่ 5

4. ผลการวิจัย

ในหัวข้อนี้ผู้วิจัยได้ทดสอบการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยด้วย LSI อัลกอริทึมนาอีฟเบย์และอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้ว โดยนำข้อมูลเว็บไซต์ท่องเที่ยวประเทศไทยจากผลการสืบค้นในเว็บไซต์กูเกิ้ลที่ได้จากระบบจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยด้วยอัลกอริทึมนาอีฟเบย์ [7] มาเป็นข้อมูลในการทดสอบจำนวน 475 เว็บไซต์ ซึ่งจัดให้อยู่ใน 6 หมวดหมู่ตามออนโทโลจีท่องเที่ยว [24] คือ สถานที่ท่องเที่ยว ที่พัก ร้านอาหาร ร้านขายของฝาก ร้านหนึ่งตำบลหนึ่งผลิตภัณฑ์และเทศกาล โดยผู้วิจัยใช้ค่าความคล้ายคลึงมากที่สุดในการจัดหมวดหมู่เว็บไซต์ด้วย LSI และหาค่า P(NB) ในแต่ละหมวดหมู่เพื่อให้ได้ค่า P(NB) ต่ำสุดและค่า P(NB) สูงสุดมาใช้ในการวิเคราะห์การจัดหมวดหมู่เว็บไซต์ด้วยอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้ว ดังแสดงตัวอย่างผลการคำนวณด้วย LSI ในตารางที่ 4 ผลการคำนวณด้วยอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้วในตารางที่ 5 และผลการจัดหมวดหมู่เว็บไซต์ด้วยอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้วในตารางที่ 6

ตารางที่ 4. ตัวอย่างผลการคำนวณด้วย LSI

URL	Similarity Value					
	Att	Acc	Res	Sou	OTOP	Even
1. http://travel.edtguide.com	0.90	0.24	0.89	-0.13	-0.21	0.24
2. http://10tis.com	0.92	0.60	0.36	-0.16	-0.18	-0.21
3. http://moohin.com	0.36	0.67	0.42	-0.16	-0.21	-0.13
4. http://www.taluitamtawan.com	0.21	-0.24	-0.13	-0.18	0.20	0.59
5. http://www.sawadee.co.th/isan/ festivals.html	0.08	0.18	0.15	0.18	-0.13	0.63
6. http://www.painaidii.com	0.61	0.24	0.95	-0.13	-0.16	-0.07
7. http://www.thaihoteltravel.com/th/ phitsanulok_thailand/local_product.htm	0.81	0.11	0.21	-0.21	-0.03	0.77
8. http://www.thailandexhibition.com	0.13	0.57	0.39	0.48	0.24	0.51
9. http://otopphitsanulok.wordpress.com	-0.21	0.48	0.49	0.59	0.55	0.31
10. http://www.hudasouvenirs.com	0.85	0.65	0.86	-0.21	-0.14	0.14
:						
475. http://tourthai.ekstepza.ws	0.31	-0.21	-0.21	-0.18	-0.16	0.05

จากตารางที่ 4 อัลกอริทึม LSI จะใช้ค่าความคล้ายคลึงที่มีค่ามากที่สุดในการระบุหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทย ตัวอย่างเช่น เว็บไซต์ที่ 1 http://travel.edtguide.com จัดอยู่ในหมวดสถานที่ท่องเที่ยว (Att) มีค่าความคล้ายคลึงมากที่สุดเท่ากับ 0.90 เว็บไซต์ที่ 3 http://moohin.com จัดอยู่ในหมวดร้านอาหาร (Res) มีค่าความคล้ายคลึงมากที่สุดเท่ากับ 0.67 เป็นต้น

ตารางที่ 5. ตัวอย่างผลการคำนวณด้วยอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้ว

URL	ค่า P(NB) ในแต่ละหมวดหมู่					
	Att	Acc	Res	Sou	OTOP	Even
1. http://travel.edtguide.com	-4.12	-0.53	-6.00	0.00	0.00	-1.61
2. http://10tis.com	-4.95	-0.44	-1.75	0.00	0.00	-0.11
3. http://moohin.com	-8.64	-0.76	-2.51	0.00	0.00	-0.98
4. http://www.taluitamtawan.com	-1.42	-0.26	-0.74	0.00	0.00	0.00
5. http://www.sawadee.co.th/isan/ festivals.html	-4.00	-0.29	-1.51	-0.02	-0.15	-4.21
6. http://www.painaidii.com	-5.94	-0.71	-8.25	0.00	0.00	-2.63
7. http://www.thaihoteltravel.com/th/ phitsanulok_thailand/local_product.htm	-0.54	-0.28	-0.16	-1.12	-0.03	-0.09

ตารางที่ 5.(ต่อ) ตัวอย่างผลการคำนวณด้วยอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้ว

URL	ค่า P(NB) ในแต่ละหมวดหมู่					
	Att	Acc	Res	Sou	OTOP	Even
8. http://www.thailandexhibition.com	-2.28	-0.10	-0.75	0.00	0.00	-2.32
9. http://otopphitsanulok.wordpress.com	-0.41	0.00	-0.13	0.00	-1.17	0.00
10. http://www.hudasouvenirs.com	-0.21	0.00	0.00	-0.47	0.00	0.00
:						
:						
475. http://tourthai.ekstepza.ws	-4.09	-0.05	0.00	0.00	0.00	0.00

เมื่อคำนวณด้วยอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงเสร็จแล้ว ซึ่งจะได้ค่า P(NB) ในแต่ละหมวดหมู่ดังแสดงในตารางที่ 5 และนำค่า P(NB) ในแต่ละหมวดหมู่มาเปรียบเทียบกับตารางที่ 2 จะได้ผลลัพธ์ในการจัดหมวดหมู่เว็บไซต์ดังแสดงในตารางที่ 6

ตารางที่ 6. ตัวอย่างผลการจัดหมวดหมู่เว็บไซต์ด้วยอัลกอริทึมนาอีฟเบย์ที่ปรับปรุง

URL	Att	Acc	Res	Sou	OTOP	Even
1. http://travel.edtguide.com	✓	✓	✓			✓
2. http://10tis.com	✓	✓	✓			✓
3. http://moohin.com	✓	✓	✓			✓
4. http://www.taluitamtawan.com	✓	✓	✓			
5. http://www.sawadee.co.th/isan/ festivals.html	✓	✓	✓	✓	✓	✓
6. http://www.painaidii.com	✓	✓	✓			✓
7. http://www.thaihoteltravel.com/th/ phitsanulok_thailand/local_product.htm	✓	✓	✓	✓	✓	✓
8. http://www.thailandexhibition.com	✓	✓	✓			✓
9. http://otopphitsanulok.wordpress.com	✓		✓		✓	
10. http://www.hudasouvenirs.com	✓			✓		
:						
:						
475. http://tourthai.ekstepza.ws	✓					

ผลลัพธ์ที่ได้จากการทดสอบเว็บไซต์จำนวน 475 เว็บไซต์จะเห็นว่าเว็บไซต์ต่าง ๆ สามารถถูกจัดให้อยู่ได้มากกว่า 1 หมวดหมู่ (ตารางที่ 6) โดยมีเว็บไซต์ http://www.sawadee.co.th/isan/festivals.html และ http://www.thaihoteltravel.com/th/phitsanulok_thailand/local_product.htm สามารถจัดให้หมวดหมู่ได้มากที่สุด 6 หมวดหมู่ คือ สถานที่ท่องเที่ยว (Att) ที่พัก (Acc) ร้านอาหาร (Res) ร้านขายของฝาก (Sou) ร้านสินค้าหนึ่งตำบลหนึ่งผลิตภัณฑ์ (OTOP) และเทศกาล (Even) และมีเว็บไซต์ที่ถูกจัด

หมวดหมู่น้อยที่สุด 1 หมวดหมู่ คือ http://tourthai.ekstepza.ws โดยถูกจัดให้อยู่ในหมวดสถานที่ท่องเที่ยว (Att)

หลังจากการวิเคราะห์เว็บไซต์เพื่อจัดหมวดหมู่ท่องเที่ยวแล้ว โดยใช้คำในการหาความถี่ของคำที่ปรากฏในแต่ละเว็บไซต์จำนวน 167 คำเพื่อจัดหมวดหมู่ 6 หมวดหมู่ คือ สถานที่ท่องเที่ยว (Att) ที่พัก (Acc) ร้านอาหาร (Res) ร้านขายของฝาก (Sou) ร้านสินค้าหนึ่งตำบลหนึ่งผลิตภัณฑ์ (OTOP) และเทศกาล (Even) ผู้วิจัยยังได้ใช้ F-Measure [25] มาทำการวัดประสิทธิภาพของการวิเคราะห์เว็บไซต์ท่องเที่ยวประเทศไทยตามสมการ 12 และแสดงการเปรียบเทียบประสิทธิภาพของการวิเคราะห์เว็บไซต์ ดังแสดงในตารางที่ 7

$$\text{F-Measure} = 2 * \left(\frac{P*R}{P+R} \right) \quad (12)$$

P คือ True Positive/(True Positive + False Positive)
R คือ True Positive/(True Positive + False Negative)
True Positive คือ เว็บไซต์ที่อยู่ในหมวดหมู่และโปรแกรมทำนายว่าอยู่ในหมวดหมู่นั้น
False Positive คือ เว็บไซต์ที่ไม่อยู่ในหมวดหมู่และโปรแกรมทำนายว่าอยู่ในหมวดหมู่นั้น
False Negative คือ เว็บไซต์ที่อยู่ในหมวดหมู่และโปรแกรมทำนายว่าไม่อยู่ในหมวดหมู่นั้น

ตารางที่ 7. เปรียบเทียบประสิทธิภาพของการวิเคราะห์เว็บไซต์

หมวดหมู่	Latent Semantic Indexing			Naïve Bayes			Modified Naïve Bayes		
	P	R	F	P	R	F	P	R	F
Att	81.82	57.80	67.74	70.31	88.46	78.35	100	95.63	97.69
Acc	87.76	54.43	67.19	94.68	63.12	75.74	100	100	100
Res	71.70	96.20	82.16	83.15	91.36	87.06	100	100	100
Sou	81.82	66.67	73.47	55.56	55.56	55.56	100	80.83	88.00
OTOP	80.00	63.16	70.59	100	58.82	74.07	100	95.24	97.44
Even	41.38	38.71	40.00	65.38	62.96	64.15	100	93.46	96.34
เฉลี่ย	**74.08**	**62.83**	**66.86**	**78.18**	**70.05**	**72.49**	**100**	**94.19**	**96.58**

จากตารางที่ 7 แสดงการเปรียบเทียบประสิทธิภาพของการวิเคราะห์เว็บไซต์ในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยด้วย LSI อัลกอริทึมนาอีฟเบย์และอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้ว จากนั้นพิจารณาในแต่ละเว็บแต่ละอัลกอริทึมตรงกับหมวดหมู่ที่ควรเป็นหรือไม่แล้วคำนวณหาค่าความแม่นยำ ค่าความระลึกและค่า F-Measure โดย LSI มีค่าความแม่นยำเฉลี่ยเท่ากับ 74.08% ค่าความระลึกเฉลี่ยเท่ากับ 62.83% ค่า F-Measure เท่ากับ 66.86 เนื่องจากบางเว็บไซต์ที่นำเสนอเนื้อหาเน้นเฉพาะหมวดใดหมวดหนึ่ง แต่พบว่ามีเนื้อหาจำนวนมากที่เกี่ยวข้องกับอีกหลายหมวดหมู่ทำให้เกิดข้อผิดพลาดในการวิเคราะห์หมวดหมู่เว็บไซต์ สำหรับอัลกอริทึมนาอีฟเบย์มีค่าความแม่นยำเฉลี่ยเท่ากับ 78.18% ค่าความ

ระลึกเฉลี่ยเท่ากับ 70.05% ค่า F-Measure เท่ากับ 72.49% เนื่องจากผลของการคำนวณด้วยอัลกอริทึมนาอีฟเบย์ที่ให้ค่าความน่าจะเป็นเพียงค่าเดียวและความถี่ของคำที่ค้นพบในหมวดหมู่ที่ควรจะเป็นน้อยทำให้การจัดหมวดหมู่ไม่ถูกต้อง เช่น หมวดร้านขายของฝาก (Sou) และหมวดร้านหนึ่งตำบลหนึ่งผลิตภัณฑ์ (OTOP) มีค่าความถูกต้อง 55.56% และ 65.38% ตามลำดับและอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้วมีค่าความแม่นยำเฉลี่ยเท่ากับ 100% เป็นผลมาจากการปรับปรุงอัลกอริทึมนาอีฟเบย์ให้แต่ละเว็บไซต์สามารถถูกจัดได้มากกว่า 1 หมวดหมู่ รวมถึงการรวบรวมคำที่เพียงพอต่อการจำแนกหมวดหมู่เว็บไซต์ทำให้การจัดหมวดหมู่เว็บไซต์มีประสิทธิภาพมากขึ้น เมื่อพิจารณาจากค่าเฉลี่ยรวมของ F-Measure จะเห็นได้ว่าวิเคราะห์การจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยด้วยอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้วมีประสิทธิภาพมากที่สุด ดังนั้นจะเห็นว่าอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้วมีความเหมาะสมในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยมากที่สุด อีกทั้งผู้วิจัยยังได้ใช้ความถี่ของคำที่ปรากฏในเว็บไซต์มาพิจารณาลำดับความสำคัญในการแสดงผลการวิเคราะห์การจัดหมวดหมู่เว็บไซต์เพื่อให้ผู้ใช้ได้ข้อมูลท่องเที่ยวที่เป็นประโยชน์มากที่สุดอีกด้วย

5. สรุป

ในงานวิจัยนี้ผู้วิจัยได้นำเสนอผลของการปรับปรุงอัลกอริทึมนาอีฟเบย์ในการวิเคราะห์เว็บไซต์ท่องเที่ยวประเทศไทยเพื่อเพิ่มประสิทธิภาพในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทย โดยใช้ข้อมูลท่องเที่ยวจากเว็บไดเร็กทอรี่ truehits เป็นชุดข้อมูลเรียนรู้ในการหาค่าความน่าจะเป็นต่ำสุดและสูงสุดในแต่ละหมวดหมู่แล้วใช้ผลลัพธ์ของการสืบค้นในเว็บไซต์กูเกิ้ลจากระบบจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยด้วยอัลกอริทึมนาอีฟเบย์ [7] เป็นชุดข้อมูลการทดสอบเริ่มจาก LSI อัลกอริทึมนาอีฟเบย์ และอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้ว จากนั้นเปรียบเทียบประสิทธิภาพด้วย F-Measure ของทั้ง 3 อัลกอริทึม พบว่าอัลกอริทึมนาอีฟเบย์ที่ปรับปรุงแล้วมีประสิทธิภาพดีที่สุด โดยมีค่า F-Measure เท่ากับ 96.58%

ในอนาคตผู้วิจัยจะนำผลลัพธ์จากการวิเคราะห์หมวดหมู่เว็บไซต์ท่องเที่ยวด้วยอัลกอริทึมนาอีฟ-เบย์ที่ปรับปรุงแล้วไปใช้ในระบบแนะนำข้อมูลท่องเที่ยวประเทศไทยและเชื่อมโยงกับออนโทโลยีเชิงเวลาในการนำเสนอข้อมูลท่องเที่ยวในช่วงเวลานั้นหรือในบริเวณใกล้เคียงกับสถานที่ท่องท่องเที่ยวที่ต้องการไป ซึ่งจะทำให้ผู้ใช้ไม่พลาดเทศกาลสำคัญ (ปฏิทินจันทรคติ) ไม่พลาดอาหารรสเด็ด (เวลาเปิดปิดร้านอาหาร) ไม่ไปเที่ยวตามช่วงเวลาที่ไม่ปลอดภัย (ปีนเขาในฤดูฝน) และทำให้เกิดความสะดวกต่อผู้ใช้อีกด้วย

เอกสารอ้างอิง (References)

[1] Zelikovitz, S. and Kogan, M., 2006. Using web searches on important words to create background sets for LSI classification. In: G. Sutcliffe & R. Goebel (Eds.), FLAIRS Conference, 598-603.

[2] Qi, X. and Davison, B.D., 2009. Web page classification: Features and algorithms. *ACM Computing Surveys*, 41(2), 1-31.

[3] Zhang, W., Yoshida, T. and Tang, X., 2008. TF-IDF, LSI and Multi-Word in Information Retrieval and Text Categorization. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 108-113.

[4] Lucia, A.D., Risi, M. and Tortora, G., 2007. Clustering Algorithms and Latent Semantic Indexing to Identify Similar Pages in Web Applications. Proceedings of the 9th IEEE International Symposium on Web Site Evolution, IEEE CS Press, 65-72.

[5] Khan, A., Baharudin, B., Lee, L. H. and Khan, K., 2010. A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, 1(1), 4-20.

[6] คมคิด ชัชราภรณ์, ธรา อั่งสกุล และจิติมนต์ อั่งสกุล, 2554. แบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยวโดยใช้เทคนิคการเรียนรู้ของเครื่อง. *วารสารเทคโนโลยีสุรนารี*, 6(2), 35-58. [Komkid Chatcharaporn, Thara Angskun and Jitimon Angskun, 2011. Tourist Attraction Categorization Models using Machine Learning Techniques. *Suranaree Journal of Social Science*, 6(2), 35-58. (in Thai)]

[7] นฤพนธ์ พนาวงศ์ และจักรกฤษณ์ เสน่ห์ นมะหุต, 2556. ระบบจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยด้วยอัลกอริทึมนาอีฟเบย์. การประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 9, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพมหานคร, 83-89. [Naruepon Panawong and Chakkrit Snae Namahoot, 2013. Thailand Tourism Web Clustering System using Naive Bayes Algorithm. The 9th National Conference on Computing and Information Technology, King Mongkut's University of Technology North Bangkok, Bangkok, 83-89. (in Thai)]

[8] Gore, S. and Pitale, R., 2013. Web Mining: An Approach towards Information Retrieval from Web with Cloud Mining. *International Journal of Computer Science and Applications*, 6(2), 190-196.

[9] เกรียงกมล คำมา และจักรกฤษณ์ เสน่ห์ นมะหุต, 2555. ระบบวิเคราะห์เว็บไซต์อนาจารด้วยกลุ่มคำเชิงความหมายคำอนาจารใน HTML Tags. การประชุมวิชาการระดับชาติ วิทยาศาสตร์วิจัย ครั้งที่ 4, มหาวิทยาลัยนเรศวร, พิษณุโลก, 17-21. [Kriangkamon Khumma and Chakkrit Snae Namahoot, 2012. Web pornography filtering by semantic network for inappropriate word analysis in HTML Tags. The 4th Science Research Conference, Naresuan University, Phitsanulok, 17-21. (in Thai)]

[10] เกรียงกมล คำมา และจักรกฤษณ์ เสน่ห์ นมะหุต, (2556). ระบบคัดกรองเว็บไซต์อนาจารด้วยเทคนิคการวิเคราะห์เว็บไซต์. การประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 9, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพมหานคร, 315-321. [Kriangkamon Khumma and Chakkrit Snae Namahoot, 2013. Pornographic Website Filtering System by Website Analysis Technique. The 9th National Conference on

Computing and Information Technology, King Mongkut's University of Technology North Bangkok, Bangkok, 315-321. (in Thai)]

[11] Chau, M. and Chen, H., 2008. A machine learning approach to web page filtering using content and structure analysis. *Decision Support Systems*, 44, 482-494.

[12] Nogueira, T.M., Rezende, S.O. and Camargo, H.A., 2010. On The Use of Fuzzy Rules to Text Document Classification. International Conference on Hybrid Intelligent Systems, 19-24.

[13] Sureshkumar, K.K., Umadevi, M., Elango, N.M., 2013. Divisive Clustering method using Naïve Bayes Algorithm for Text Categorization. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(4), 1747-1753.

[14] Thongkrau, T. and Lalitrojwong, P., 2010. Classifying Instances into Lexical Ontology Concepts Using Latent Semantic Analysis. The 2nd International Conference on Computer and Automation Engineering, 66-70.

[15] Thongbai, R., Viriyapant, K., 2013. A Development Questions Answering Student Services System Using Latent Semantic Indexing. The 9th National Conference on Computing and Information Technology, 822-827.

[16] Bhat, V., Oates, T., Shanbhag, V. and Nicholas, C., 2004. Finding aliases on the web using latent semantic analysis. Data & Knowledge Engineering, 49, 129-143.

[17] กนกรัตน์ จิรสัจจานุกูล และณัฐธนนท์ หงส์วริทธิ์ธร, 2555. ระบบการวิเคราะห์ข้อคิดเห็นของผู้เชี่ยวชาญด้วยการวิเคราะห์ ความหมายแอบแฝงและการจัดกลุ่มข้อความ. *วารสารเทคโนโลยีสารสนเทศ*, 8(1), 1-12. [Kanokrat Jirasatjanukul and Nuttanont Hongwarittorn, 2012. Analyze System of Expert's Opinions using Latent Semantic Analysis and Text Clustering Technique. *Information Technology Journal*, 8(1), 1-12. (in Thai)]

[18] Thorleuchter, D. and Poel, D.V.D., 2013. Technology classification with latent semantic indexing, *Expert Systems with Application*, 40, 1786-1795.

[19] กองบรรณาธิการโปรวิชั่น, 2552. 10000+Web รวมสุดยอดเว็บไซท์. [Provision, 2009. 10000+Web (in Thai)]

[20] Eckart, C., Young, G., 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), 211-218.

[21] Salton, G., and McGill, M.J., 1983. Introduction to Modern Information Retrieval. New York: McGraw-Hill.

[22] Nivet Chirawichitchai and Narin Panawas, 2012. Sentiment Classification Using Machine Learning Techniques. [online] Available at: http://www.east.spu.ac.th/~narin.pa/Sentiment_ Classification.pdf [Accessed 20 October 2012].

[23] Patil, A.S. and Pawar, B.V., 2012. Automated Classification of Web Sites using Naïve Bayesian Algorithm. Proceedings of the International MultiConference of Engineers and Computer Scientists 2012, 519-523.

[24] นฤพนธ์ พนาวงศ์ และจักรกฤษณ์ เสน่ห์ นมะหุต, 2553. ระบบค้นหาสถานที่ท่องเที่ยวในประเทศไทยด้วยหลักการออนโทโลจีและเนมแมทชิง. *Journal of Information Science and Technology*, 1(2), 60-69. [Naruepon Panawong and Chakkrit Snae, 2010. Search System for Attractions in Thailand with Ontology and Name Matching. *Journal of Information Science and Technology*, 1(2), 60-69. (in Thai)]

[25] นฤพนธ์ พนาวงศ์ และจักรกฤษณ์ เสน่ห์ นมะหุต, 2556. การวิเคราะห์ประสิทธิภาพของระบบสืบค้นข้อมูลออนโทโลจีท่องเที่ยวด้วยอัลกอริทึม ISG และ Name Variation Matching. วารสารวิทยาศาสตร์ มหาวิทยาลัยนเรศวร, 9(2), 47-64. [Naruepon Panawong and Chakkrit Snae Namahoot, 2013. Performance Analysis of an Ontology-Based Tourism Information System with ISG Algorithm and Name Variation Matching. *NU Science Journal*, 9(2), 47-64. (in Thai)]