

ระบบจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยด้วยอัลกอริทึมนาอ์เบย์

Thailand Tourism Web Clustering System using Naïve Bayes Algorithm

นฤพนธ์ พนาวงศ์ (Naruepon Panawong)¹ และจักรกฤษณ์ เสน่ห์ นมะหุด (Chakkrit Snae Namahoot)²

ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนครสวรรค์

¹inaruepon.p@gmail.com, ²chakkrits@nu.ac.th

บทคัดย่อ

การท่องเที่ยวถือเป็นอุตสาหกรรมที่มีความสำคัญต่อการพัฒนาเศรษฐกิจภายในประเทศเป็นอย่างมากจากการรณรงค์กระตุ้นและส่งเสริมการท่องเที่ยวจากภาครัฐ เมื่อนักท่องเที่ยวต้องการข้อมูลท่องเที่ยวส่วนใหญ่ใช้วิธีการค้นหาผ่านเว็บเสิร์ชเอนจินอย่างกูเกิ้ล แต่ผลลัพธ์ของการค้นหาที่ได้มีจำนวนมากมีทั้งข้อมูลที่ไม่ตรงกับความต้องการหรือมีข้อมูลที่ไม่เกี่ยวข้องกับท่องเที่ยวปะปนอยู่และไม่ถูกจัดหมวดหมู่ท่องเที่ยวรวมถึงข้อมูลท่องเที่ยวไม่ถูกนำเสนอในเว็บไซต์เดียว ทำให้เสียเวลาในการเข้าถึงข้อมูลจากหลายเว็บไซต์เพื่อให้ได้ข้อมูลท่องเที่ยวตามต้องการทั้งหมด คือ เที่ยวที่ไหน กินที่ไหน พักที่ไหน ร้านขายของฝากที่ไหน ดังนั้นผู้วิจัยจึงใช้อัลกอริทึมนาอ์เบย์ในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยตามออนโทโลยีที่ได้ออกแบบไว้เพื่อช่วยในการเข้าถึงข้อมูลท่องเที่ยวเป็นไปอย่างสะดวกรวดเร็วภายในเว็บไซต์เดียวและใช้ F-Measure วัดประสิทธิภาพของระบบที่ได้พัฒนาด้วยอัลกอริทึมนาอ์เบย์นี้ ซึ่งพบว่ามีค่าความแม่นยำเท่ากับ 72.60 % ค่าความระลึกเท่ากับ 70.99 % และค่า F-Measure เท่ากับ 71.61 %

คำสำคัญ: จัดหมวดหมู่ นาอ์เบย์ ออนโทโลยี การจัดการข้อมูล การจัดการความรู้ การเข้าถึงข้อมูล

Abstract

Through encouragement and advancement of national and international tourism campaigns by the Thai government the tourism industry plays a very important role in developing Thailand's economy. To extract tourism information most travelers rely on such Web search engines as Google, but the results are mixed with several pieces of information. Moreover, most of them are not categorized in tourism area so that users have to

waste time accessing data on multiple Websites to collect the information that meets their needs, for example, attractive places, restaurants, hotels, and souvenir shops. To deal with this problem, we have used naive Bayes to cluster tourism Websites in Thailand with the help of an ontology. Consequently, the system can provide a faster way to gather tourism information in one pass. F-Measure has been used to validate the system, and the results show 72.60% for precision, 70.99% for recall, and 71.61% for F-measure.

Keyword: Classification, Naïve Bayes, Ontology, Information Management System, Knowledge Management, Information Retrieval

1. บทนำ

การท่องเที่ยวถือเป็นอุตสาหกรรมบริการที่ทำรายได้ให้แก่ประเทศต่าง ๆ เป็นจำนวนมาก รวมทั้งประเทศไทย โดยมีบทบาทอย่างมากต่อการพัฒนาเศรษฐกิจภายในประเทศ เนื่องจากเป็นอุตสาหกรรมที่มีการเติบโตอย่างต่อเนื่อง รวมทั้งได้มีการพัฒนาแหล่งท่องเที่ยว วิธีการจัดการการท่องเที่ยวให้ขยายตัวมากขึ้น ซึ่งเป็นปัจจัยที่สามารถสร้างเศรษฐกิจหมุนเวียนและเป็นการกระจายรายได้ไปสู่ท้องถิ่นภายในประเทศเป็นอย่างดี อีกทั้งจากการแข่งขันจากประเทศเพื่อนบ้าน การเปิดประชาคมเศรษฐกิจอาเซียน ซึ่งการท่องเที่ยวแห่งประเทศไทยได้จัดกิจกรรมส่งเสริมและประชาสัมพันธ์การท่องเที่ยวของประเทศไทยให้กับนักท่องเที่ยวไทยและชาวต่างประเทศอย่างต่อเนื่องผ่านทางเว็บไซต์ แผ่นพับ โทรทัศน์ และสื่อต่าง ๆ อีกมากมาย ซึ่งในปี 2555 การท่องเที่ยวแห่งประเทศไทยได้จัดโครงการ “เที่ยวหัวใจใหม่ เมืองไทยยั่งยืน” โดยเจาะตลาดภายในประเทศและต่างประเทศ พร้อมมุ่งรักษาคุณค่าทางการท่องเที่ยวและความประทับใจ เพื่อเป็นการ

กระตุ้นเศรษฐกิจและเร่งการฟื้นตัวหลังจากเกิดปัญหาอุทกภัยภายในประเทศและมุ่งหวังสร้างรายได้เพิ่มขึ้นอีก 9% [1] อีกทั้งการท่องเที่ยวภาคกลาง เปิดตัวโครงการ “เที่ยวภาคกลางทั้งปีไม่มีเบื่อ” เพื่อเป็นการต้อนรับประชาคมเศรษฐกิจอาเซียนหรือ AEC ที่กำลังมาถึงในปี 2558 [2]

ปัจจุบันมีเว็บไซต์จำนวนมากที่ให้บริการค้นหาข้อมูลท่องเที่ยวหรือระบบที่ช่วยให้ข้อมูลเกี่ยวกับการท่องเที่ยว แต่การที่นักท่องเที่ยวไปเที่ยวยังสถานที่แห่งใดแห่งหนึ่งจำเป็นต้องรู้สถานที่ท่องเที่ยว ที่พัก ร้านอาหาร ร้านขายของฝากหรือของที่ระลึก ร้านขายสินค้าหนึ่งตำบลหนึ่งผลิตภัณฑ์ เทศกาลท่องเที่ยวประจำท้องถิ่นนั้น ๆ ซึ่งนักท่องเที่ยวส่วนใหญ่มักค้นหาผ่านทางเสิร์ชเอนจินอย่างกูเกิ้ล แต่ผลลัพธ์ที่ได้จากการค้นหาอาจสร้างความลำบากให้แก่นักท่องเที่ยวได้ เช่น ข้อมูลที่ได้มีจำนวนมาก ข้อมูลที่ได้ไม่ตรงกับความต้องการทั้งหมด ข้อมูลที่ได้ไม่ถูกจัดหมวดหมู่ไว้อย่างเหมาะสม ทำให้เสียเวลาในการคัดเลือกข้อมูลท่องเที่ยวเพื่อให้ได้ข้อมูลท่องเที่ยวที่ตรงกับความต้องการทั้งหมด

ดังนั้นผู้วิจัยจึงใช้อัลกอริทึมนาอิวเบย์ในระบบจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยและวัดประสิทธิภาพของอัลกอริทึมนาอิวเบย์ด้วย F-Measure โดยอธิบายงานวิจัยที่เกี่ยวข้องในหัวข้อที่ 2 วิธีการดำเนินงานวิจัยในหัวข้อที่ 3 ผลการดำเนินงานในหัวข้อที่ 4 และสรุปในหัวข้อที่ 5

2. งานวิจัยที่เกี่ยวข้อง

ในการจำแนกหมวดหมู่เว็บเพจมีความแตกต่างจากการจำแนกหมวดหมู่เอกสารทั่วไปที่มักมีโครงสร้าง เนื้อหา และคุณลักษณะที่ชัดเจนซึ่งแตกต่างจากเว็บเพจ การเรียนรู้คุณลักษณะเพื่อจำแนกหมวดหมู่เว็บเพจนั้นอาจอาศัยเนื้อหาที่ปรากฏบนหน้าเว็บเพจ แต่อาจพบว่ามีสิ่งรบกวนต่าง ๆ เช่น โฆษณา การเชื่อมโยง ขยะทางอิเล็กทรอนิกส์ หรือแม้แต่โครงสร้างของเว็บเพจที่ไม่เป็นไปตามมาตรฐานอันเนื่องมาจากความผิดพลาดระหว่างการประมวลผล [3]

ปัจจุบันมีนักวิจัยต่าง ๆ นำเสนอแนวความคิดในการจัดหมวดหมู่เว็บไซต์หรือจัดกลุ่มของเอกสารออนไลน์ด้วยเทคนิคและอัลกอริทึมต่าง ๆ ดังนี้ [4] นำเสนองานวิจัยเกี่ยวกับอัลกอริทึมการเรียนรู้ของเครื่องในการการจัดกลุ่มเอกสาร

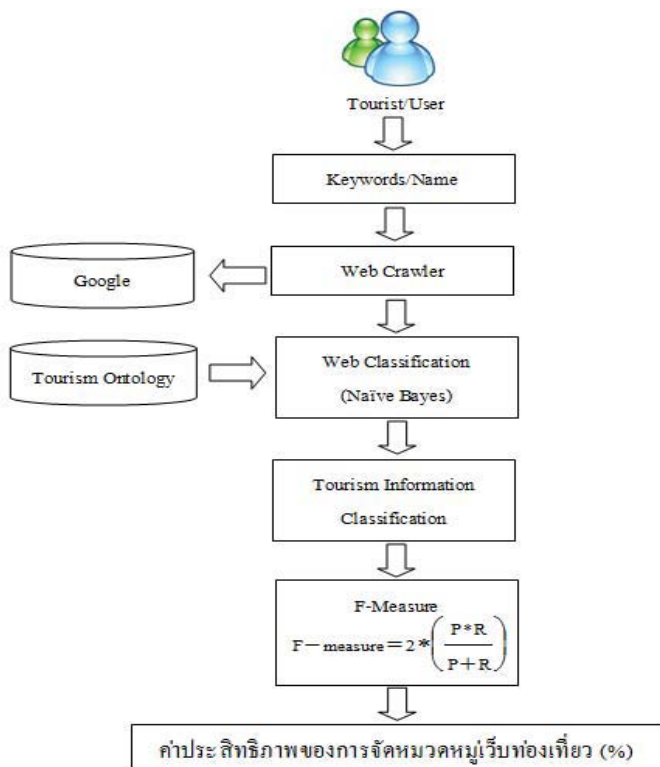
ออนไลน์จำพวกข่าวออนไลน์ บล็อก อีเมลและห้องสมุดดิจิทัล รวมถึงใช้การประมวลผลภาษาธรรมชาติมาใช้ในการกระบวนการตัดคำหยุด การแทนด้วยรากศัพท์ของคำ พบว่าอัลกอริทึมนาอิวเบย์ทำงานได้ดีในการกรองสแปม จัดกลุ่มอีเมล ตัวเลขและข้อมูลที่เป็นข้อความ ต้องการจำนวนข้อมูลที่ใช้เรียนรู้บ่อย ง่ายต่อการพัฒนาเมื่อเปรียบเทียบกับอัลกอริทึมอื่น ๆ แต่การประมวลผลภาษาธรรมชาติมีความยุ่งยาก ซับซ้อน เสียเวลาในการประมวลผล ซึ่ง [5] ก็ได้เปรียบเทียบการทำงานของอัลกอริทึม LBK J48 Naïve Bayes และ OneR เพื่อจัดกลุ่มของเอกสารด้วยโปรแกรมเวก้า ซึ่งพบว่าอัลกอริทึมนาอิวเบย์มีผลลัพธ์ที่ดีที่สุด แต่ผลลัพธ์ที่ได้ไม่สามารถแสดงค่าประสิทธิภาพของการเปรียบเทียบอย่างทันที ต่อมา [6] ได้นำเสนอเทคนิคการจำแนกความสนใจของนักท่องเที่ยวต่อการท่องเที่ยวในประเทศไทย โดยใช้การทำเหมืองข้อความและออนโทโลยีเพื่อจำแนกชุดข้อความความสนใจซึ่งจำแนกออกเป็น 5 โดเมน คือ สถานที่ท่องเที่ยว ที่พัก การเดินทาง กิจกรรมและเทศกาล ซึ่งเก็บข้อความสนใจจากหัวข้อกระทู้และข้อความ แสดงความคิดเห็น ในกระทู้จากเว็บไซต์ tripadvisor.com และสามารถนำมาประยุกต์ใช้ในงานวิจัยนี้ได้ จากนั้น [7] ได้ออกแบบและพัฒนาฐานความรู้ออนโทโลยีสำหรับวิเคราะห์ข่าวออนไลน์โดยอัตโนมัติ โดยใช้ข่าวออนไลน์ภาษาไทยที่เกี่ยวข้องกับพื้นที่จังหวัดชายแดนใต้จำนวน 200 ข่าวและจัดกลุ่มคำสำคัญที่มีความสอดคล้องกัน และประเมินคำสำคัญที่มีผลต่อการวิเคราะห์ข่าวออนไลน์ด้วยวิธีการหาค่าความถี่ของคำที่ปรากฏในเนื้อหาข่าวออนไลน์ โดยสามารถนำแนวคิดมาประยุกต์ใช้ในงานวิจัยนี้เป็นอย่างดี ซึ่ง [8] ได้จำแนกความคิดเห็นโดยใช้เทคนิคการเรียนรู้ของเครื่องด้วยอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน ค้นไม่ตัดสินใจ นาอิวเบย์และเคเนียร์สเนเบอร์ พบว่าอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีนและนาอิวเบย์มีประสิทธิภาพในการจำแนกข้อมูลความคิดเห็นของผู้ชมที่เขียนจากความรู้สึกและอารมณ์ของผู้ใช้ที่มีต่อการใช้สินค้าและบริการนั้น ๆ และเมื่อแทนค่าคุณลักษณะด้วยค่าความจริงและลดคุณลักษณะด้วย Information Gain พบว่าอัลกอริทึมนาอิวเบย์มีประสิทธิภาพดีที่สุดและ [9] ได้สร้างแบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยวโดยใช้เทคนิคการเรียนรู้ของเครื่อง โดยใช้ข้อมูล

สถานที่ท่องเที่ยวจากเว็บไซต์โลนลี่ แพลนเน็ตมาจัดกลุ่มสถานที่ท่องเที่ยวจำนวน 5 กลุ่มคือ สถานที่พัก สถานที่ท่องเที่ยวยามค่ำคืน ร้านอาหาร แหล่งจับจ่ายสินค้า สถานที่เยี่ยมชม แล้ววัดประสิทธิภาพของแบบจำลองทั้งหมด 4 เทคนิคคือ ต้นไม้ตัดสินใจ นาอ็ฟเบย์ ซัพพอร์ทเวกเตอร์แมชชีน พบว่า นาอ็ฟเบย์ให้ค่าความถูกต้องสูงที่สุด แต่อย่างไรก็ตามแบบจำลองนี้นำเสนอเฉพาะข้อมูลสถานที่ท่องเที่ยวที่เป็นภาษาอังกฤษเท่านั้น

จากงานวิจัยที่เกี่ยวข้องทำให้ผู้วิจัยเลือกใช้อัลกอริทึม นาอ็ฟเบย์มาจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทย เนื่องจากเป็นอัลกอริทึมที่ง่าย ไม่ซับซ้อนและให้ค่าความถูกต้องมากที่สุด ดังได้กล่าวไว้ข้างต้นแล้ว อีกทั้งยังสามารถนำไปประยุกต์ใช้กับเว็บไซต์ที่แสดงผลภาษาไทยได้ง่ายอีกด้วย

3. วิธีการดำเนินการวิจัย

ในงานวิจัยนี้ได้ออกแบบสถาปัตยกรรมของระบบ (ภาพที่ 1) และทดสอบการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยด้วยอัลกอริทึม นาอ็ฟเบย์ ดังรายละเอียดในหัวข้อที่ 3.1 และ 3.2



ภาพที่ 1 : แสดงสถาปัตยกรรมของระบบ

จากภาพที่ 1 แสดงสถาปัตยกรรมของระบบซึ่งมีการทำงานดังต่อไปนี้

1. ผู้ใช้ป้อนชื่อจังหวัดหรือชื่อสถานที่ท่องเที่ยวหรือข้อมูลท่องเที่ยว
2. นำชื่อจังหวัดหรือชื่อสถานที่ท่องเที่ยวหรือข้อมูลท่องเที่ยวในข้อที่ 1 ไปสืบค้นข้อมูลในเว็บไซต์ที่ถูกรวบรวมด้วยเทคนิค Web Crawler
3. นำผลลัพธ์ที่ได้จากข้อที่ 2 มาจัดหมวดหมู่เว็บไซต์ท่องเที่ยวด้วยอัลกอริทึม นาอ็ฟเบย์ (สมการที่ 1) โดยค่าที่ใช้ในการจัดหมวดหมู่ได้มาจากออนโทโลยีท่องเที่ยว และใช้ค่าน้อยที่สุดของ C_{MAP} เป็นข้อสรุปว่าอยู่ในหมวดหมู่ใด
4. แสดงผลของการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวและแสดงค่าประสิทธิภาพการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวด้วย F-Measure

ในงานวิจัยนี้ได้นำแนวคิดมาจาก [10] ที่ใช้อัลกอริทึม นาอ็ฟเบย์ในการจัดหมวดหมู่เว็บไซต์มาปรับปรุงเพื่อจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยจำนวน 6 หมวดหมู่คือ สถานที่ท่องเที่ยว ที่พัก ร้านอาหาร ร้านขายของฝาก ร้านหนึ่งตำบลหนึ่งผลิตภัณฑ์และเทศกาล โดยมีขั้นตอนในการทำงาน 2 ส่วนคือ กระบวนการเรียนรู้และกระบวนการทดสอบจัดหมวดหมู่เว็บไซต์ท่องเที่ยว ดังรายละเอียดต่อไปนี้

3.1 การเรียนรู้

เป็นขั้นตอนที่ใช้เป็นข้อมูลเพื่อนำไปทดสอบหาความน่าจะเป็นของเว็บไซต์เหล่านั้นควรจัดอยู่ในหมวดหมู่ใด ในงานวิจัยนี้ใช้ฐานข้อมูลท่องเที่ยวประเทศไทยจากเว็บ truehits จากนั้นคัดเลือกมาเฉพาะ 6 หมวดหมู่ที่ต้องการแล้วนำมาใช้ในการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทย ดังแสดงในตารางที่ 1

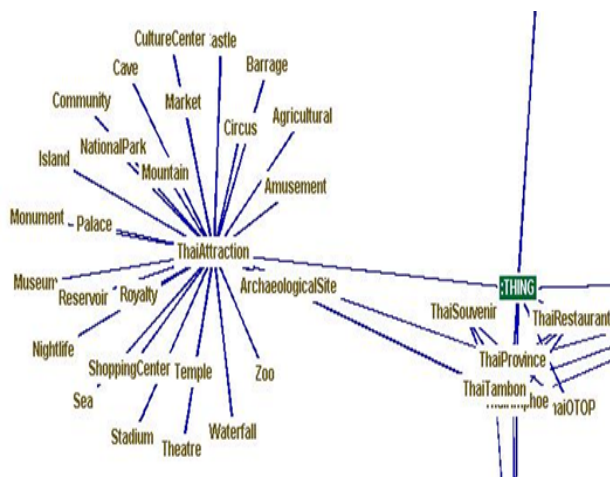
3.2 การทดสอบจัดหมวดหมู่เว็บไซต์

ในงานวิจัยนี้ได้ทำการทดสอบการจัดหมวดหมู่เว็บไซต์ท่องเที่ยว โดยใช้ผู้ป้อนคำค้นหา “กินไหน นอนไหน เที่ยวไหนดี ในพิษณุโลก” จากนั้นนำคำค้นหาไปสืบค้นในเว็บไซต์ที่ถูกรวบรวมแล้วนำผลลัพธ์ที่ได้มาจัดหมวดหมู่ตามออนโทโลยีท่องเที่ยว [11] ได้แก่ คลาสสถานที่ท่องเที่ยว คลาสที่พัก คลาสร้านอาหาร คลาสร้านขายของฝาก คลาสร้านหนึ่งตำบลหนึ่งผลิตภัณฑ์และคลาสเทศกาล ดังแสดงตัวอย่างในภาพที่ 2 รวมถึงใช้ค่าที่ได้จาก

ออนโทโลยีท่องเที่ยวมาหาความถี่ที่ปรากฏในเอกสารแต่ละเว็บไซต์

ตารางที่ 1: แสดงจำนวนเว็บไซต์และตัวอย่างคำที่ใช้ในการจัดหมวดหมู่

หมวดหมู่ (G)	ตัวอย่างคำที่ใช้	จำนวนเว็บไซต์
1	เที่ยว ภูเขา น้ำตก พิพิธภัณฑ์ ถ้ำ	233
2	ที่พัก โรงแรม รีสอร์ท โฮมสเตย์	200
3	ร้านอาหาร	318
4	ของฝาก	54
5	หนึ่งตำบลหนึ่งผลิตภัณฑ์	88
6	เทศกาล สถาปัตยกรรม กีฬา	155
รวม		1,048



ภาพที่ 2 : แสดงตัวอย่างออนโทโลยีท่องเที่ยว

จากนั้นเมื่อได้ความถี่ของคำ โปรแกรมก็ทำการคำนวณหาความน่าจะเป็นที่เอกสารเว็บควรจัดอยู่ในหมวดหมู่ใด โดยคำนวณทีละหมวดหมู่ จากนั้นนำผลของการคำนวณทั้ง 6 หมวดหมู่มารวมและเลือกผลของการคำนวณหมวดหมู่ที่มีค่าน้อยที่สุดเป็นข้อสรุปในการจัดหมวดหมู่เว็บไซต์ด้วยอัลกอริทึมแนฟเบย์ ดังแสดงในสมการที่ 1

$$C_{MAP} = \underset{c \in C}{\operatorname{argmin}}(C_i) \quad (1)$$

$$C_i = \log(P(c)) + \sum_{i=1}^n \log(P(t_i | c)) \quad (2)$$

เมื่อ d คือ เอกสารเว็บเพจ

C_i คือ ผลการคำนวณในแต่ละกลุ่ม $\{c_1, c_2, c_3, \dots, c_n\}$

C คือ กลุ่มที่ต้องการจัดหมวดหมู่ $\{c_1, c_2, c_3, \dots, c_n\}$

$P(c)$ คือ ความน่าจะเป็นของกลุ่มหรือหมวดหมู่คำนวณได้จากสมการที่ 3

$$P(c) = \frac{N_c}{N} \dots \dots \dots (3)$$

เมื่อ N_c คือ จำนวนเอกสารเรียนรู้ทั้งหมดในกลุ่ม (c)

N คือ จำนวนเอกสารที่ใช้ในการเรียนรู้ทั้งหมด

$P(t | c)$ คือความน่าจะเป็นของเอกสารในแต่ละหมวดหมู่คำนวณได้จากสมการที่ 4

$$P(t | c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)} \dots \dots \dots (4)$$

เมื่อ T_{ct} คือ ความถี่ของคำ (t) ในเอกสาร d จากกลุ่ม c

$\sum_{t' \in V} T_{ct'}$ คือ ผลรวมความถี่ของคำในเอกสาร d จากกลุ่ม c

เนื่องจากผลของการคำนวณจากสมการที่ 2 ได้ค่าติดลบในทุกหมวดหมู่และจากความถี่ของคำค้นหาไปในทางกลุ่มที่มีผลการคำนวณติดลบมากที่สุด ดังนั้นผู้วิจัยจึงเลือกใช้ค่าน้อยที่สุดมาใช้ในการพิจารณาเว็บไซต์นั้นควรจัดอยู่ในหมวดหมู่ใด

4. ผลการดำเนินงาน

ในหัวข้อนี้แสดงผลการทำงานของกรอบการออกแบบหน้าจอที่ผู้ใช้งานต้องป้อนคำค้นหาข้อมูลท่องเที่ยว จากนั้นระบบก็แสดงข้อมูลท่องเที่ยวตามหมวดหมู่ด้วยอัลกอริทึมแนฟเบย์ โดยผู้วิจัยได้พัฒนาเป็นเว็บที่มีการใช้งานง่ายคล้ายเว็บเสิร์ชเอนจินทั่วไปด้วยภาษา JSP ดังแสดงในภาพที่ 3 และผลของการคำนวณด้วยอัลกอริทึมแนฟเบย์ดังแสดงตัวอย่างในตารางที่ 2 และการวัดประสิทธิภาพของการจัดหมวดหมู่เว็บดังแสดงในตารางที่ 3

ค้นหา ค้นหา

ข้อมูลท่องเที่ยวตามหมวดหมู่ (ผลการค้นหาจากเฉลี่ย 1,110,000 รายการ)

- สถานที่ท่องเที่ยว www.painaidii.com
- ที่พัก www.thammachartresorthotel.com
- ร้านอาหาร www.moohin.com/food/026
- ร้านขายของฝาก www.hotelsthailand.com/north/phitsanulok/souvenir.html
- ร้านหนึ่งตำบลหนึ่งผลิตภัณฑ์ otopphitsanulok.wordpress.com
- เทศกาล www.ryt9.com/s/iqry/1594149

ภาพที่ 3 : แสดงหน้าจอตัวอย่างการจัดหมวดหมู่ข้อมูลท่องเที่ยว

ตารางที่ 2: แสดงตัวอย่างผลของการคำนวณด้วยอัลกอริทึมเบย์

Naïve Bayes						
กลุ่ม	1	2	3	4	5	6
URL 1	www.painaidii.com สถานที่ท่องเที่ยว - ไปไหนดี อะไรดี ที่ไหนดี					
	-31.59	-9.15	-0.52	-1.29	-1.08	-27.35
URL 2	www.thammachartresorthotel.com โรงแรม ที่พัก พิชญโลก					
	-8.48	-9.53	-0.52	-1.29	-1.08	-0.83
URL 3	www.moohin.com/food/026 รายชื่อร้านอาหารจังหวัดพิจญโลก					
	-18.54	-3.73	-20.52	-1.29	-1.08	-0.83
URL 4	www.hotelsthailand.com/north/phitsanulok/souvenir.html พิจญโลก - ของฝาก ของที่ระลึก					
	-28.62	-13.18	-0.52	-42.29	-1.08	-16.91
URL 5	otopphitsanulok.wordpress.com หนึ่งตำบล หนึ่งผลิตภัณฑ์ (OTOP) จังหวัดพิจญโลก					
	-0.65	-0.72	-0.52	-1.29	-1.68	-0.83
URL 6	http://www.ryt9.com/s/iqry/1594149 พิจญโลก จัดเทศกาลโคมไฟ ประจำปี 22 - 24 ก.พ.56					
	-24.65	-0.72	-0.52	-1.29	-1.08	-28.59

จากตารางที่ 2 แสดงตัวอย่างผลของการคำนวณด้วยอัลกอริทึมเบย์ในแต่ละหมวด โดย URL 1 เป็นเว็บไซต์ที่ให้รายละเอียดเกี่ยวกับสถานที่ท่องเที่ยวและถูกจัดหมวดหมู่ในกลุ่มของสถานที่ท่องเที่ยว (กลุ่มที่ 1) เนื่องจากในกลุ่มนี้มีค่าน้อยที่สุด (-31.59) ตามการคำนวณในสมการที่ 2

URL 2 เป็นเว็บไซต์โรงแรม ที่พัก ในจังหวัดพิจญโลกและถูกจัดหมวดหมู่ในกลุ่มของที่พัก (กลุ่มที่ 2) เนื่องจากในกลุ่มนี้มีค่าน้อยที่สุด (-9.53) ตามการคำนวณในสมการที่ 2

URL 3 เป็นเว็บไซต์รายชื่อร้านอาหารในจังหวัดพิจญโลกและถูกจัดหมวดหมู่ในกลุ่มร้านอาหาร (กลุ่มที่ 3) เนื่องจากในกลุ่มนี้มีค่าน้อยที่สุด (-20.52) ตามการคำนวณในสมการที่ 2

URL 4 เป็นเว็บไซต์ของฝาก ของที่ระลึกที่มีในจังหวัดพิจญโลกและถูกจัดหมวดหมู่ในกลุ่มของร้านขายของฝาก (กลุ่มที่ 4) เนื่องจากในกลุ่มนี้มีค่าน้อยที่สุด (-42.29) ตามการคำนวณในสมการที่ 2

URL 5 เป็นเว็บไซต์หนึ่งตำบล หนึ่งผลิตภัณฑ์ในจังหวัดพิจญโลกและถูกจัดหมวดหมู่ในกลุ่มของร้านหนึ่งตำบลหนึ่งผลิตภัณฑ์ (กลุ่มที่ 5) เนื่องจากในกลุ่มนี้มีค่าน้อยที่สุด (-1.68) ตามการคำนวณในสมการที่ 2

URL 6 เป็นเว็บไซต์เทศกาลในจังหวัดพิจญโลกและถูกจัดหมวดหมู่ในกลุ่มของเทศกาล (กลุ่มที่ 6) เนื่องจากในกลุ่มนี้มีค่าน้อยที่สุด (-28.59) ตามการคำนวณในสมการที่ 2

หลังจากจัดหมวดหมู่เว็บไซต์แล้วผู้วิจัยยังได้ใช้ F-Measure [12] มาทำการวัดประสิทธิภาพของการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยด้วยอัลกอริทึมเบย์ ตามสมการที่ 5 และแสดงค่าประสิทธิภาพของการจัดหมวดหมู่เว็บไซต์ท่องเที่ยวประเทศไทยด้วยอัลกอริทึมเบย์ ดังแสดงในตารางที่ 3

$$F - \text{measure} = 2 * \left(\frac{P * R}{P + R} \right) \quad (5)$$

เมื่อ 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 3: แสดงค่าประสิทธิภาพการจัดหมวดหมู่เว็บไซต์

G	TP	FP	FN	P	R	F-measure
1	153	48	29	76.12 %	84.07 %	79.90 %
2	114	29	27	79.72 %	80.85 %	80.28 %
3	57	18	24	76.00 %	70.37 %	73.08 %
4	17	9	10	65.38 %	62.96 %	64.15 %
5	11	7	6	61.11 %	64.71 %	62.86 %
6	17	5	10	77.27 %	62.96 %	69.39 %
รวม				72.60 %	70.99 %	71.61 %

จากตารางที่ 3 ใช้ผลลัพธ์จากการสืบค้นในเว็บไซต์กูเกิ้ล ทั้งหมด 475 เว็บไซต์ จากนั้นพิจารณาในแต่ละเว็บไซต์ตรงกับหมวดหมู่ที่ควรเป็นหรือไม่แล้วคำนวณหาค่าความแม่นยำ ค่าความระลึกและค่า F-Measure โดยมีค่าความแม่นยำอยู่ในเกณฑ์ที่ดี เนื่องจากเว็บไซต์ส่วนใหญ่ใช้คำที่งานวิจัยนี้รองรับ แต่ก็ยังมีบางเว็บไซต์ที่จัดหมวดหมู่ไม่ถูกต้อง เนื่องจากมีการใช้คำที่ในงานวิจัยนี้ยังไม่ครอบคลุมและบางเว็บไซต์เป็นไฟล์รูปภาพหรือไฟล์ PDF

5. สรุป

งานวิจัยนี้ได้นำเสนอระบบจัดหมวดหมู่เว็บไซต์ท่องเที่ยวในประเทศไทยด้วยอัลกอริทึมนาอูฟเบย์เพื่อนำเสนอข้อมูลท่องเที่ยวให้เป็นหมวดหมู่ทำให้สะดวกต่อการเข้าถึงข้อมูลท่องเที่ยวในเว็บไซต์เดียว โดยใช้ข้อมูลท่องเที่ยวจากเว็บไคเร็กทอรี่ truehits เป็นชุดข้อมูลเรียนรู้และใช้ผลลัพธ์ของการสืบค้นในเว็บไซต์กูเกิ้ลเป็นชุดข้อมูลทดสอบการจัดหมวดหมู่เว็บไซต์ท่องเที่ยว โดยใช้คำและหมวดหมู่จากออนโทโลยีท่องเที่ยว จากนั้นทำการวัดประสิทธิภาพของอัลกอริทึมนาอูฟเบย์โดยมีค่าความแม่นยำเท่ากับ 72.60 % ถือว่าอยู่ในเกณฑ์ดี ค่าความระลึก เท่ากับ 70.99 % และค่า F-Measure เท่ากับ 71.61 %

แต่อย่างไรก็ตามคำที่ใช้ในการจัดหมวดหมู่เว็บนั้นยังไม่ครอบคลุมเนื้อหาหรือเว็บไซต์ท่องเที่ยวทั้งหมด เช่น ที่กิน ที่ดื่ม หมายถึง ร้านอาหาร ที่นอน ที่หลับ หมายถึง ที่พัก เป็นต้น ดังนั้นในอนาคตผู้วิจัยจะใช้คำเหมือนหรือคล้ายคลึงจากพจนานุกรม เช่น ร้านอาหาร อาจเรียกว่า ภัตตาคาร โรงอาหาร ที่ขายอาหาร เหล่า เป็นต้น เพื่อให้สื่อความหมายถึงคำเหล่านั้น

รวมถึงพัฒนาอัลกอริทึมในการตรวจสอบความถี่ของคำที่พบในเว็บไซต์ท่องเที่ยวแล้วนำค่าไปเข้ากระบวนการวิเคราะห์ความเกี่ยวข้องของเว็บไซต์ต่าง ๆ ที่บ่งบอกถึงลักษณะเฉพาะของเว็บไซต์ท่องเที่ยว เพื่อเพิ่มประสิทธิภาพของการจัดหมวดหมู่เว็บไซต์และใช้อัลกอริทึมการทำเหมืองข้อมูลอื่น ๆ มาเปรียบเทียบประสิทธิภาพเพื่อเพิ่มความน่าเชื่อถือของงานวิจัยมากยิ่งขึ้น

เอกสารอ้างอิง

- [1] ผู้จัดการออนไลน์ [ออนไลน์]. เข้าถึงจาก: <http://www.manager.go.th/Travel/ViewNews.aspx?NewsID=9540000080652>, สืบค้น 10 พฤศจิกายน 2555.
- [2] ผู้จัดการออนไลน์ [ออนไลน์]. เข้าถึงจาก: <http://www.manager.go.th/Travel/ViewNews.aspx?NewsID=9560000006234>, สืบค้น 11 กุมภาพันธ์ 2556.
- [3] กมลาสน์ อุคมล้ำเลิศ และ อานนท์ รุ่งสว่าง, “การคัดแยกหมวดหมู่เว็บเพจภาษาไทยอัตโนมัติด้วยนาอูฟเบย์”, *การประชุมทางวิชาการของมหาวิทยาลัยเกษตรศาสตร์ ครั้งที่ 48: สาขาสถาปัตยกรรมศาสตร์และวิศวกรรมศาสตร์* 3-5 ก.พ. 2553 หน้า 310-317.
- [4] A. Khan, B. Baharudin, L. H. Lee and K. Khan, “A Review of Machine Learning Algorithms for Text-Documents Classification”, *Journal of Advances in Information Technology*, Vol. 1, No. 1, pp. 4-20, 2010.
- [5] T. M. Nogueira, S. O. Rezende and H. A. Camargo, “On The Use of Fuzzy Rules to Text Document Classification”, *International Conference on Hybrid Intelligent Systems*, pp. 19-24, 2010.
- [6] นิศา สถาพรพวงนา, “การใช้การทำเหมืองข้อความและออนโทโลยีในการจำแนกความสนใจของนักท่องเที่ยวต่อการท่องเที่ยวในประเทศไทย”, *วิทยานิพนธ์ วท.ม.*, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 2553.
- [7] วิชุดา โชติรัตน์, ศุภดิ บุญรอด และ ศศิมา งาม วิเชียร, “การพัฒนาฐานความรู้ออนโทโลยีสำหรับวิเคราะห์ข่าวออนไลน์โดยอัตโนมัติ”, *วารสารเทคโนโลยีสารสนเทศ* ปีที่ 7 ฉบับที่ 14 ก.ค. – ธ.ค. 2554 หน้า 13-18.
- [8] นิเวศ จิระวิจิตรชัย และ นรินทร์ พนาวาส, “การจำแนกความคิดเห็นโดยใช้เทคนิคการเรียนรู้ของเครื่อง”, สืบค้นเมื่อ 20 ตุลาคม 2555, จาก http://www.east.spu.ac.th/~narin.pa/Sentiment_Classification.pdf.
- [9] คมกิต ชัยราภรณ์, ธรา อังสกุล และ จิตินันต์ อังสกุล, “แบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยวโดยใช้เทคนิคการเรียนรู้ของ

เครื่อง”, วารสารเทคโนโลยีสุรนารี ปีที่ 6 ฉบับที่ 2 ข.ค. 2554 หน้า 35-58.

- [10] A. S. Patil and B. V. Pawar, “Automated Classification of Web Sites using Naïve Bayesian Algorithm”, *Proceedings of the International MultiConference of Engineers and Computer Scientists* 2012, pp. 519-523, 2012.
- [11] นฤพนธ์ พนาวงศ์ และ จักรกฤษณ์ เสน่ห์, “ระบบค้นหาสถานที่ท่องเที่ยวในประเทศไทยด้วยหลักการออนโทโลยีและเนมแมทชิง”, *Journal of Information Science and Technology*, Vol. 1, No. 2, July - December 2010 หน้า 60-69.
- [12] P. Christen and K. Goiser, “Quality and complexity measures for data linkage and deduplication”, *Quality Measures in Data Mining, Studies in Computational Intelligence*, Vol. 43, pp. 127-151, 2007.