แบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยวโดยใช้เทคนิคการเรียนรู้ของเครื่อง

Tourist AttractionCategorizationModelsusing Machine Learning Techniques

คมคิด ชัชราภรณ์*, ธรา อั่งสกุล และจิติมนต์ อั่งสกุล
Komkid Chatcharaporn*, Thara Angskun and Jitimon Angskun

สาขาวิชาเทคโนโลยีสารสนเทศ สำนักวิชาเทคโนโลยีสังคม มหาวิทยาลัยเทคโนโลยีสุรนารี

Abstract

The selection of tourism attractions is an important activity for travelers. The Internet is a significant data source to support travelers for choosing attractions. Nevertheless, there are many tourist attractions on the Internet which are uncategorized. Hence, this work has focused on the development of tourist attraction categorization models using machine learning techniques.It starts with the development of a module for collecting tourist attraction data from the Internet automatically, and then these data are filtered by selecting suitable features for model construction. The feature selection is divided into two cases: 1) The feature selection from tourist attraction name only; and 2) The feature selection from tourist attraction name and description. Then the tourist attraction data of selected features are processed through four machine learning techniques to construct the categorization models. The four machine learning techniques comprise Decision Trees, Naïve Bayes, Support Vector Machine and k-Nearest Neighbor. The comparison of the model performance is considered from accuracy, precision, recall and time period taken to build the models. The experimental results reveal that the models constructed from the features of the second case can provide average accuracy, precision and recall values higher than that of the first case. The machine learning technique that is the most effective for tourist attraction categorization is Naïve Bayes which achieves with 88.70% average of accuracy at 700 features.

Keywords: Text Categorization; Tourist Attraction; Machine Learning

บทคัดย่อ

การเลือกสถานที่ท่องเที่ยวเป็นหนึ่งกิจกรรมที่สำคัญของนักท่องเที่ยว อินเทอร์เน็ตได้กลายมาเป็นแหล่งข้อมูลสำคัญที่ช่วยในการเลือกสถานที่ท่องเที่ยวเหล่านั้น แต่อย่างไรก็ตาม ยังมีสถานที่ท่องเที่ยวจำนวนมหาศาลที่ยังไม่ได้ถูกจัดไว้เป็นหมวดหมู่ ดังนั้นบทความนี้จึงมุ่งเน้นการพัฒนาแบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยวแบบอัตโนมัติโดยใช้เทคนิคการเรียนรู้ของเครื่อง โดยเริ่มตั้งแต่การพัฒนาโมดูลเพื่อช่วยในการรวบรวมข้อมูลสถานที่ท่องเที่ยวแบบอัตโนมัติจากแหล่งข้อมูลบนอินเทอร์เน็ตและนำข้อมูลเหล่านั้นมาคัดกรองโดยการเลือกคุณลักษณะที่เหมาะสมในการสร้างแบบจำลอง โดยการเลือกคุณลักษณะนั้นแบ่งออกเป็น 2กรณี คือ1)เลือกคุณลักษณะจากชื่อสถานที่ท่องเที่ยวเท่านั้น และ 2) เลือกคุณลักษณะจากชื่อและคำอธิบายของสถานที่ท่องเที่ยวหลังจากนั้น นำข้อมูลเฉพาะคุณลักษณะที่ถูกเลือกมาผ่านกระบวนการเรียนรู้ของเครื่อง 4 เทคนิค ได้แก่ ต้นไม้การตัดสินใจ นาอีฟเบย์สซัพพอร์ทเวกเตอร์แมชชีน และเคเนียเรสเนเบอร์เพื่อสร้างแบบจำลองการจัดหมวดหมู่ สำหรับการเปรียบเทียบประสิทธิภาพของแบบจำลองพิจารณา

* *ผู้เขียนที่ให้การติดต่อ*
E-mail address: komkid1@hotmail.com

จากค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และระยะเวลาที่ใช้ในการสร้างแบบจำลอง ผลการทดลองพบว่า แบบจำลองซึ่งสร้างจากข้อมูลตามคุณลักษณะของกรณีที่ 2 ให้ค่าความถูกต้อง ค่าความแม่นยำ และค่าความระลึกโดยเฉลี่ยสูงกว่ากรณีที่ 1ส่วนเทคนิคการเรียนรู้ของเครื่องที่มีประสิทธิภาพมากที่สุดในการจัดหมวดหมู่สถานที่ท่องเที่ยวคือ นาอีฟเบย์สซึ่งใช้คุณลักษณะจำนวน 700 คุณลักษณะโดยให้ค่าความถูกต้องเฉลี่ย 88.70%

บทนำ

การเลือกสถานที่ท่องเที่ยว (Tourist Attraction) ไม่ว่าจะเป็นร้านอาหาร โรงแรม ที่พัก สถานที่เยี่ยมชมที่มีชื่อเสียงไปจนถึงแหล่งเลือกซื้อสินค้าและของฝาก เป็นหนึ่งกิจกรรมที่นักท่องเที่ยวให้ความสำคัญเป็นอย่างยิ่งเมื่อต้องตัดสินใจเดินทางท่องเที่ยว (Huang and Bian, 2009) โดยเฉพาะอย่างยิ่งการเดินทางไปยังสถานที่ที่ไม่คุ้นเคยหรือไม่เคยไปมาก่อนนักท่องเที่ยวจำเป็นต้องหาข้อมูลมาประกอบการตัดสินใจ ซึ่งในปัจจุบันอินเทอร์เน็ตได้กลายมาเป็นแหล่งข้อมูลที่สำคัญที่นักท่องเที่ยวเข้าถึงเพื่อค้นหาและพิจารณาเลือกสถานที่ท่องเที่ยวในขอบเขตที่ตนเองสนใจว่าตรงกับความต้องการของตนหรือไม่ อีกทั้งการเข้าถึงข้อมูลบนอินเทอร์เน็ตไม่จำกัดการเข้าถึงเฉพาะที่อีกต่อไป เนื่องจากอินเทอร์เน็ตผ่านโทรศัพท์มือถือ (Mobile Internet) เป็นอีกหนึ่งเทคโนโลยีสำคัญที่ช่วยให้นักท่องเที่ยวสามารถเข้าถึงข้อมูลสถานที่ท่องเที่ยวได้ทุกที่ทุกเวลา (Egger and Buhalis, 2008)

ในปัจจุบันมีเว็บไซต์และบริการค้นหาข้อมูลออนไลน์หลายแหล่งที่อำนวยความสะดวกในการค้นหาข้อมูลสถานที่ท่องเที่ยวบนอินเทอร์เน็ต การค้นหาข้อมูลสถานที่ท่องเที่ยวสำหรับนักท่องเที่ยวในสถานที่ไม่รู้จักมาก่อน ข้อมูลหรือผลลัพธ์ที่ได้จากการค้นหาอาจสร้างความลำบากให้แก่นักท่องเที่ยวได้ เช่น ข้อมูลสถานที่ท่องเที่ยวที่ไม่ได้มีการระบุหมวดหมู่ไว้ว่าเป็นสถานที่ท่องเที่ยวหมวดหมู่ใดหรือสถานที่ท่องเที่ยวชื่อเดียวกันแต่แหล่งข้อมูลคนละแหล่งมีการระบุหมวดหมู่ของสถานที่ที่แตกต่างกันออกไป นอกจากนี้ชื่อสถานที่ท่องเที่ยวที่เป็นชื่อเฉพาะ ยกตัวอย่างเช่นสถานที่ท่องเที่ยวในประเทศไทย อย่างสยามพารากอน (Siam Paragon)หรือ มาบุญครอง (Mahboonkrong) ชื่อสถานที่อาจไม่ได้สื่อโดยตรงว่าสถานที่เหล่านี้เป็นแหล่งจับจ่ายสินค้าประเด็นเหล่านี้ส่งผลให้นักท่องเที่ยวจำเป็นต้องค้นหาข้อมูลเกี่ยวกับสถานที่นั้น ๆ เพิ่มเติมว่าแท้จริงแล้วสถานที่ท่องเที่ยวนั้นจัดอยู่ในหมวดหมู่อะไร ตรงกับความต้องการของตนหรือไม่

จากการทบทวนวรรณกรรมพบว่ามีงานวิจัยหลายงานที่มุ่งแก้ปัญหาเรื่องของการระบุหรือจัดหมวดหมู่ของสถานที่ท่องเที่ยวโดยส่วนใหญ่เป็นการประยุกต์ใช้เทคนิคของการทำเหมืองข้อมูล (Data Mining) เข้ามาช่วยรับมือกับข้อมูลจำนวนมหาศาลนั่นคือ เทคนิคการจัดหมวดหมู่ (Classification) เมื่อนำเทคนิคนี้มาใช้กับเนื้อหาบนอินเทอร์เน็ตซึ่งอยู่ในรูปแบบข้อความ (Text) เรียกเทคนิคการจัดหมวดหมู่กับข้อมูลประเภทข้อความนี้ว่า การจัดหมวดหมู่ข้อความ (Text Classification) โดยหน้าที่หลักของการจัดหมวดหมู่ข้อความ คือ การระบุประเภทของเอกสารหรือข้อความแบบอัตโนมัติโดยอิงจากตัวเนื้อหาในเอกสารหรือข้อความนั้น ๆ (Miller and Nguyen, 2005; Sebastiani, 2005) ดังนั้นเทคนิคนี้จึงมีการนำศาสตร์ด้านอื่น ๆ เข้ามาผสมด้วย เช่น การค้นคืนสารสนเทศ (Information Retrieval)

และการประมวลผลด้วยภาษาธรรมชาติที่มาจากศาสตร์ทางด้านปัญญาประดิษฐ์ (Artificial Intelligence) เป็นต้น

การนำการจัดหมวดหมู่ข้อความมาช่วยในการจัดหมวดหมู่ของสถานที่ท่องเที่ยว จะเป็นการค้นหารูปแบบและความสัมพันธ์ในชุดข้อมูลโดยมีการนำศาสตร์ด้านปัญญาประดิษฐ์ ได้แก่ การเรียนรู้ของเครื่อง (Machine Learning) มาช่วยเรียนรู้และสร้างแบบจำลองเพื่อทำนายข้อมูลของสถานที่ท่องเที่ยวที่เข้ามาใหม่ว่าควรจัดอยู่ในหมวดหมู่ใด สำหรับงานวิจัยที่เกี่ยวข้องกับการจัดหมวดหมู่สถานที่ท่องเที่ยวที่พบในปัจจุบันมีดังนี้ งานวิจัยแรกมีการนำเมตาดาต้า (Meta Data) ของรูปภาพสถานที่ท่องเที่ยวซึ่งอยู่ในรูปแบบข้อความ ที่ได้จากเว็บไซต์ผู้ให้บริการอัพโหลดรูปภาพฟลิคเกอร์ (Flickr) การจำแนกและระบุสถานที่ท่องเที่ยวโดยใช้การทำเหมืองข้อมูลเพื่อสกัดเอาข้อมูลสถานที่และระยะเวลาที่นักท่องเที่ยวอัพโหลดรูปมาวิเคราะห์และแนะนำตารางการเดินทางภายในหนึ่งวันว่าสามารถไปเที่ยวสถานที่ใดได้บ้าง (Popescu, Grefenstette and Moëllic, 2009a) งานวิจัยถัดมาได้จำแนกหมวดหมู่สถานที่แบ่งออกเป็น 2 หมวดหมู่ คือ สถานที่ที่มีอยู่ ซึ่งเป็นสถานที่ที่ระบุไว้ในฐานข้อมูลของสมุดหน้าเหลืองและสถานที่ที่พบใหม่โดยใช้ข้อมูลการทำหมายเหตุบนแผนที่ (Map Annotation) จากผู้ใช้บริการแผนที่ทั่วไปบนเว็บไซต์ให้บริการของไมโครซอฟต์ งานวิจัยนี้ใช้เทคนิคการทำเหมืองข้อมูลแบบจัดกลุ่ม (Clustering) ร่วมกับเทคนิคการประมวลผลด้วยภาษาธรรมชาติ (Mummidi and Krumm, 2008) และถัดมาเป็นงานวิจัยที่มุ่งเน้นการจัดทำพจนานุกรมทางด้านภูมิศาสตร์ซึ่งมีการจัดหมวดหมู่พวกวัตถุทางด้านภูมิศาสตร์ เช่น ภูเขา แม่น้ำ เมือง โบราณสถาน ฯลฯ โดยพิจารณาถึงข้อมูลของสถานที่ต่าง ๆ ประกอบ ซึ่งข้อมูลที่นำมาใช้ในงานนี้มาจากแหล่งข้อมูลออนไลน์อย่างฟลิคเกอร์และวิกิพีเดีย (Wikipedia) การสกัดและจัดหมวดหมู่ข้อมูลหมวดหมู่พวกวัตถุทางด้านภูมิศาสตร์ใช้เทคนิคการทำเหมืองข้อมูลกับชื่อของสถานที่งานนี้ไม่ได้พิจารณาเพียงภาษาอังกฤษเท่านั้น แต่ยังมีภาษาฝรั่งเศส สเปน และอิตาลีรวมอยู่ด้วย (Popescu, Grefenstette & Bouamor, 2009b) นอกจากนี้ยังมีงานวิจัยที่มีวัตถุประสงค์เพื่อจัดหมวดหมู่สถานที่ท่องเที่ยวอีกงานที่ใช้ออนโทโลยีและการเรียนรู้ของเครื่องมาประยุกต์ใช้ โดยข้อมูลสถานที่ท่องเที่ยวจะถูกรวบรวมมาจากแหล่งข้อมูลออนไลน์ เช่น ดีบีพีเดีย (DBpedia) กูเกิ้ล แมพส์ (Google Maps) จีโอเนมส์ (GeoNames) และฐานข้อมูลที่อยู่ในท้องถิ่นนั้น ๆ มาจัดเก็บลงในออนโทโลยี (Ontology) สถานที่ที่ใดที่ผู้จัดเก็บไม่รู้แน่ชัดจะใช้การจัดหมวดหมู่ด้วยการทำเหมืองข้อมูล โดยใช้ขั้นตอนวิธีนาอีฟเบย์ส (Naïve Bayes) มาช่วย ซึ่งข้อมูลสถานที่ท่องเที่ยวที่นำมาใช้ในกระบวนการจัดหมวดหมู่ คือ ชื่อของสถานที่ท่องเที่ยว (Ozdikis, Orhan and Danismaz, 2011)

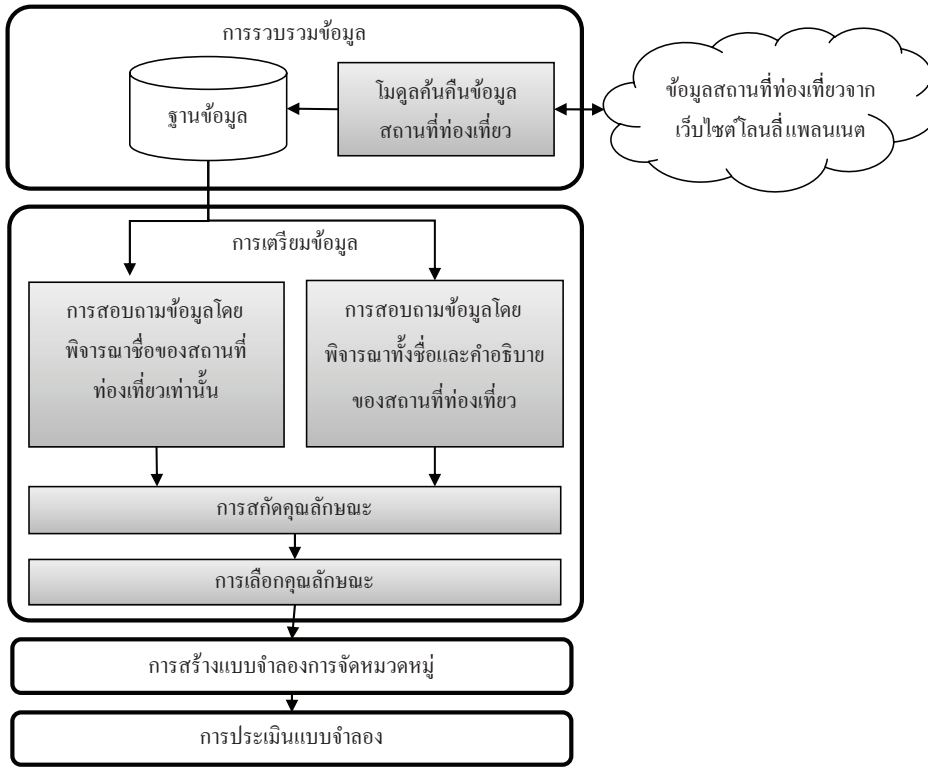
จากการทบทวนวรรณกรรมข้างต้นพบว่าข้อมูลสถานที่ท่องเที่ยวที่นำมาสร้างแบบจำลองนั้นมาจากหลายแหล่ง และการนำชื่อของสถานที่ท่องเที่ยวมาใช้ในการพิจารณาจัดหมวดหมู่ของสถานที่ในงานของป็อบสคิว (Popescu et al., 2009b) และออซดิสคิส (Ozdikis et al., 2011) ผลลัพธ์จากการทดลองสะท้อนให้เห็นถึงผลลัพธ์ที่ยังไม่เป็นที่น่าพอใจของการนำเฉพาะชื่อสถานที่ท่องเที่ยว

มาใช้กับการจัดหมวดหมู่ คือ ได้ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ที่สูงในข้อมูลสถานที่บางหมวดหมู่ (Ozdikis et al., 2011) หรือมีเฉพาะค่าความแม่นยำเท่านั้นที่สูงแต่ค่าความระลึกต่ำ (Popescu et al., 2009b; Ozdikis et al., 2011) เป็นไปได้ว่าเนื่องจากชื่อของสถานที่นั้นไม่ได้บ่งบอกถึงหมวดหมู่ของสถานที่อย่างชัดเจน ดังนั้นการจัดหมวดหมู่โดยคำนึงถึงชื่อของสถานที่ท่องเที่ยวเพียงอย่างเดียวนั้นไม่อาจสามารถทำนายถึงหมวดหมู่ของสถานที่นั้นได้อย่างมีประสิทธิภาพ

งานวิจัยนี้จึงนำเสนอวิธีการจัดหมวดหมู่สถานที่ท่องเที่ยวโดยใช้เทคนิคการเรียนรู้ของเครื่องสำหรับสร้างแบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยวจากแหล่งข้อมูลบนอินเทอร์เน็ต โดยการเปรียบเทียบวิธีการเลือกคุณลักษณะที่จะนำมาใช้สร้างตัวแบบจำลองระหว่างการเลือกคุณลักษณะจากชื่อของสถานที่ท่องเที่ยวเท่านั้น และการเลือกคุณลักษณะจากทั้งชื่อและคำอธิบายของสถานที่ท่องเที่ยวนอกจากนี้ยังมีการเปรียบเทียบประสิทธิภาพของแบบจำลองการจัดหมวดหมู่ที่ได้จากเทคนิคการเรียนรู้ของเครื่องแบบต่าง ๆรวมทั้งแสดงให้เห็นถึงความเป็นไปได้ในการรวมรวบข้อมูลของสถานที่ท่องเที่ยวออนไลน์แบบอัตโนมัติเพื่อนำมาจัดเก็บและใช้ในการสร้างแบบจำลองการจัดหมวดหมู่ต่อไป

กรอบการพัฒนาแบบจำลอง

งานวิจัยนี้มุ่งไปที่การออกแบบและพัฒนาแบบจำลองที่มีประสิทธิภาพเพื่อใช้ในการจัดหมวดหมู่สถานที่ท่องเที่ยวโดยนำข้อมูลของสถานที่ท่องเที่ยวซึ่งเป็นภาษาอังกฤษจากเว็บไซต์โลนลี่แพลนเน็ต (Lonely Planet, 2011) ซึ่งเหตผุลที่เลือกเว็บไซต์โลนลี่ แพลนเน็ต เนื่องมาจากการสำรวจแหล่งข้อมูลออนไลน์หลาย ๆ แหล่ง เว็บไซต์โลนลี่ แพลนเน็ตเป็นแหล่งที่ให้ข้อมูลของสถานที่ท่องเที่ยวที่มีทั้งชื่อ (Name) คำอธิบาย (Description) และหมวดหมู่ (Type) ของสถานที่ท่องเที่ยว ซึ่งแตกต่างจากแหล่งข้อมูลที่งานวิจัยอื่นนำมาใช้ เช่น เว็บไซต์ฟลิคเกอร์จะมีข้อมูลเพียงแค่ชื่อรูปสถานที่เท่านั้น (Flickr, 2011) เว็บไซต์วิกิพีเดียจะมีชื่อและคำอธิบาย แต่หมวดหมู่ของสถานที่ไม่ได้ถูกระบุไว้ ทำให้การรวบรวมข้อมูลจากแหล่งนี้ผู้รวบรวมต้องระบุหมวดหมู่ของข้อมูลด้วยตนเองจึงจะนำมาใช้ในการสร้างแบบจำลองได้ (Wikipedia, 2011) ส่วนข้อมูลจากเว็บไซต์กูเกิ้ล แมพส์โดยส่วนใหญ่จะมีเพียงชื่อและหมวดหมู่ของสถานที่เท่านั้นโดยหมวดหมู่ของสถานที่นั้นเจ้าของสถานที่สามารถระบุได้มากกว่าหนึ่งหมวดหมู่ส่วนคำอธิบายจะขึ้นอยู่กับเจ้าของข้อมูลสถานที่ว่าจะใส่หรือไม่ใส่ทำให้การรวบรวมข้อมูลสถานที่ท่องเที่ยวจากแหล่งนี้บางครั้งไม่ได้รับข้อมูลคำอธิบายสถานที่เสมอไป และหมวดหมู่ของสถานที่ที่หลากหลายนั้นผู้ใช้จำเป็นต้องสรุปอีกทีว่าสถานที่นั้นควรจัดอยู่ในหมวดหมู่ใด (Google, 2011) ดังนั้นจากการสำรวจแหล่งข้อมูลข้างต้นงานวิจัยนี้จึงเลือกใช้เว็บไซต์โลนลี่ แพลนเน็ตเป็นแหล่งข้อมูลในการดำเนินงานวิจัย โดยเลือกเฉพาะข้อมูลสถานที่ท่องเที่ยวที่ตั้งอยู่ในประเทศไทยมาใช้เป็นข้อมูลในการพัฒนาแบบจำลองสำหรับจัดหมวดหมู่ซึ่งขั้นตอนการพัฒนาและประเมินแบบจำลองสามารถแสดงได้ดังรูปที่ 1 ซึ่งประกอบด้วยขั้นตอนต่าง ๆ ดังต่อไปนี้



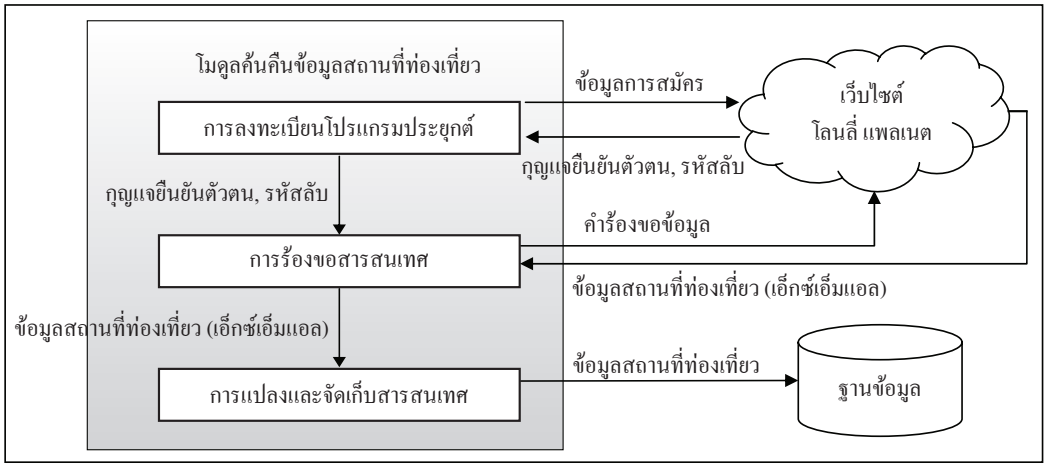
รูปที่ 1 ขั้นตอนการพัฒนาแบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยว โดยใช้เทคนิคการเรียนรู้ของเครื่อง

1. การรวบรวมข้อมูล (Data Collection) การศึกษาครั้งนี้ต้องการรวบรวมข้อมูลสถานที่ท่องเที่ยวจากเว็บไซต์โลนลี่ แพลนเน็ต (Lonely Planet) มาจัดเก็บลงในฐานข้อมูลแต่เนื่องจากการเก็บรวบรวมข้อมูลสถานที่ท่องเที่ยวจำนวนมหาศาลจากเว็บไซต์ด้วยมนุษย์ใช้เวลานาน ดังนั้นงานวิจัยนี้จึงได้พัฒนาโมดูลขึ้นมาสำหรับช่วยรวบรวมและจัดเก็บข้อมูลสถานที่ท่องเที่ยวแบบอัตโนมัติ โดยโมดูลที่พัฒนาขึ้นมานั้นสามารถร้องขอข้อมูลจากเว็บไซต์มาจัดเก็บผ่านส่วนต่อประสานโปรแกรมประยุกต์ของโลนลี่ แพลนเน็ต (Application Programming Interface:API) โดยข้อมูลของสถานที่ท่องเที่ยวที่รวบรวมมามีจำนวนทั้งสิ้น 1,606ระเบียนซึ่งเป็นข้อมูลสถานที่ท่องเที่ยวในประเทศไทย ข้อมูลทั้งหมดเป็นภาษาอังกฤษประกอบด้วยชื่อ คำบรรยาย และหมวดหมู่ของสถานที่สำหรับการแบ่งหมวดหมู่ของสถานที่ท่องเที่ยวในงานวิจัยนี้แบ่งออกเป็น 5 หมวดหมู่ (Class) ได้แก่สถานที่พัก (Hotel) สถานที่ท่องเที่ยวยามค่ำคืน (Nightlife) ร้านอาหาร (Restaurant) แหล่งจับจ่ายสินค้า (Shopping) และสถานที่เยี่ยมชม (Sigh Seeing) ซึ่งข้อมูลทั้งหมดจะถูกนำมาจัดเก็บลงฐานข้อมูลมายเอสคิวแอล (MySQL) เพื่อใช้ในกระบวนการเตรียมข้อมูลต่อไปจำนวนระเบียนข้อมูลของสถานที่ท่องเที่ยวแต่ละหมวดหมู่แสดงในตารางที่ 1

ตารางที่ 1 จำนวนระเบียนในสถานที่ท่องเที่ยวในประเทศไทยแต่ละหมวดหมู่

หมวดหมู่สถานที่ท่องเที่ยว	จำนวนระเบียน
สถานที่พัก (Hotel)	492
สถานที่ท่องเที่ยวยามค่ำคืน (Nightlife)	171
ร้านอาหาร (Restaurant)	467
แหล่งจับจ่ายสินค้า (Shopping)	223
สถานที่เยี่ยมชม (Sight Seeing)	253
รวม	**1,606**

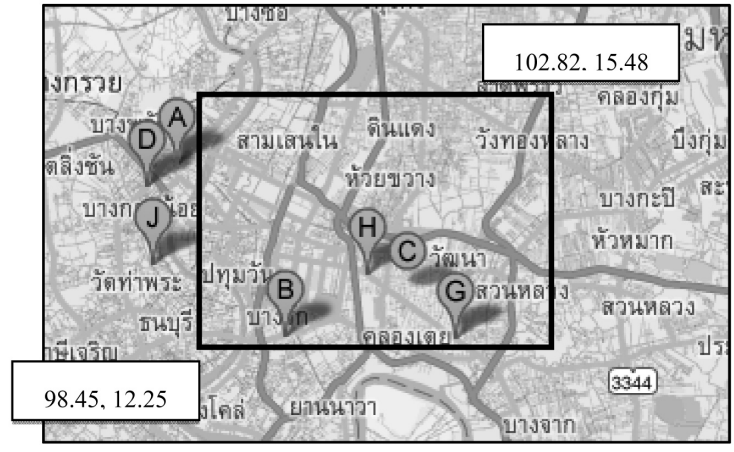
ส่วนขั้นตอนการทำงานภายในโมดูลค้นคืนข้อมูลสถานที่ท่องเที่ยวจากเว็บไซต์โลนลี่ แพลนเน็ตนั้น ประกอบไปด้วยการทำงานหลัก 3 ขั้นตอน ดังรูปที่ 2



รูปที่ 2 ขั้นตอนการทำงานของโมดูลค้นคืนข้อมูลสถานที่ท่องเที่ยว

ขั้นตอนที่ 1 การลงทะเบียนโปรแกรมประยุกต์ (Application Registration) การเข้าถึงข้อมูลสถานที่ท่องเที่ยวของเว็บโลนลี่แพลนเนตนั้นนักพัฒนาจำเป็นต้องลงทะเบียนขอใช้งานโปรแกรมประยุกต์จากทางเว็บไซต์ก่อน ซึ่งกระบวนการนี้จะกระทำเพียงครั้งเดียวเท่านั้น จากนั้นทางเว็บไซต์จะส่งมอบกุญแจยืนยันตัวตน (OAuthKey) และรหัสลับ (Secret Code) กลับมาเพื่อยืนยันความเป็นเจ้าของโปรแกรมประยุกต์คืนแก่นักพัฒนาดังนั้นในการร้องขอข้อมูลสถานที่ท่องเที่ยวทุกครั้งนักพัฒนาจำเป็นต้องส่งกุญแจยืนยันตัวตนและรหัสลับแนบไปกับคำร้องขอด้วย

ขั้นตอนที่ 2 การร้องขอสารสนเทศ (Information Query) สำหรับรูปแบบการร้องขอข้อมูลสถานที่ท่องเที่ยวของเว็บโลนลี่แพลนเนตนั้นมีหลายรูปแบบ เช่น การค้นหาจากชื่อของสถานที่ การค้นหาสถานที่โดยการกำหนดขอบเขต และการค้นหาสถานที่จากรหัส(Lonely Planet, 2010) ในงานวิจัยนี้ใช้รูปแบบการกำหนดขอบเขตของเขตพิกัดละติจูด (Latitude) กับลองจิจูด (Longitude) ที่ต้องการบนแผนที่โลกซึ่งลักษณะของขอบเขตจะอยู่ในรูปแบบของกรอบสี่เหลี่ยม (Bounding Box) ดังรูปที่ 3 นอกจากนี้คำร้องขอที่ถูกส่งไปจำเป็นต้องกำหนดหมวดหมู่ของสถานที่ท่องเที่ยวที่ต้องการควบคู่ไปด้วย



รูปที่ 3 การกำหนดขอบเขตพิกัดที่ตั้งของสถานที่ท่องเที่ยวในรูปแบบของกรอบสี่เหลี่ยม

โดยสรุปแล้วในข้อมูลการร้องขอจะประกอบไปด้วยกุญแจยืนยันตัวตนรหัสลับ ข้อมูลพิกัด และหมวดหมู่ของสถานที่ท่องเที่ยวซึ่งการส่งข้อมูลการร้องขอไปยังเครื่องแม่ข่ายของเว็บไซต์โลนลี่แพลนเนตสามารถกระทำผ่านยูอาร์แอล (Uniform Resource Locator:URL) ซึ่งมีรูปแบบดังต่อไปนี้

http://OAuthKey:Secret@api.lonelyplanet.com/api/bounding_boxes/{north},{south},{east},{west}/pois?poi_type={type-name}

ขั้นตอนที่ 3 การแปลงและจัดเก็บสารสนเทศ (Information Transformation and Storage) ผลลัพธ์ที่ได้กลับมาจากการร้องขอข้อมูลสถานที่ท่องเที่ยวจะอยู่ในรูปแบบเอกสารเอ็กซ์เอ็มแอล (Extensible Markup Language: XML) ดังรูปที่4ตัวอย่างข้อมูลภายในเอกสารประกอบด้วยข้อมูลรายละเอียดของสถานที่ เช่น รหัส ชื่อพิกัดที่ตั้งและหมวดหมู่ของสถานที่เป็นต้น

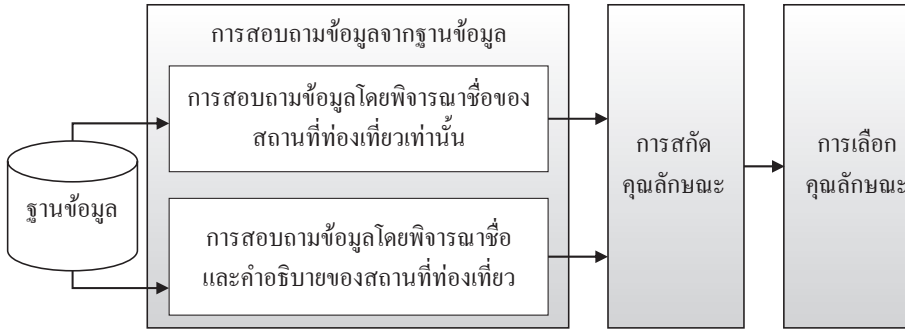
```
<pois>
<poi>
<id>367907</id>
<name>Coogee Bay Hotel</name>
<digital-latitude>-33.920940998916</digital-latitude>
<digital-longitude>151.256557703018</digital-longitude>
<poi-type>Night</poi-type>
</poi>
<poi>
<id>367711</id>
<name>New Theatre</name>
<digital-latitude>-33.9031756797962</digital-latitude>
<digital-longitude>151.179843842983</digital-longitude>
<poi-type>Night</poi-type>
</poi>
</pois>
```

รูปที่ 4 ผลการค้นคืนข้อมูลสถานที่ท่องเที่ยวในรูปแบบเอกสารเอ็กซ์เอ็มแอล

ข้อมูลสถานที่ท่องเที่ยวที่อยู่ในเอกสารรูปแบบเอ็กซ์เอ็มแอลดังกล่าวจะถูกนำมาแปลงให้อยู่ในรูปของอะเรย์ (Array) ด้วยภาษาพีเอชพี (PHP) ก่อนที่จะจัดเก็บลงฐานข้อมูล โดยเลือกข้อมูลเฉพาะชื่อ คำอธิบาย และหมวดหมู่ของสถานที่ท่องเที่ยวมาจัดเก็บลงฐานข้อมูลมายเอสคิวแอล (MySQL) ก่อนที่จะนำไปประมวลผลในขั้นตอนถัดไป

2. การเตรียมข้อมูล (Data Preprocessing) ถัดจากการรวบรวมข้อมูลเป็นขั้นตอนการนำข้อมูลสถานที่ท่องเที่ยวในฐานข้อมูลมาประมวลผลด้วยเทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing Technique) และแปลงข้อมูลให้อยู่ในรูปแบบของเมตริกซ์เอกสาร (Document Matrix) ก่อนนำไปใช้สร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่อง (Machine Learning Techniques) สำหรับรายละเอียดในขั้นตอนการเตรียมข้อมูลนี้ประกอบไปด้วยขั้นตอนย่อย ๆ ที่ปรากฏดังรูปที่ 5

2.1 การสอบถามข้อมูลจากฐานข้อมูล (Data Query from Database) เป็นการสอบถามข้อมูลสถานที่ท่องเที่ยวที่จัดเก็บไว้ในฐานข้อมูลออกมาใช้โดยแบ่งการสอบถามข้อมูลออกเป็น 2 วิธี คือ 1) การสอบถามข้อมูลโดยพิจารณาชื่อของสถานที่เท่านั้น และ 2) การสอบถามข้อมูลโดยพิจารณาทั้งชื่อและคำอธิบายของสถานที่ท่องเที่ยวซึ่งข้อมูลสถานที่ท่องเที่ยวที่ได้จากการสอบถามจากทั้ง 2 วิธีจะถูกนำมาประมวลผลด้วยเทคนิคการประมวลผลภาษาธรรมชาติเพื่อเลือกคุณลักษณะที่เหมาะสมก่อนนำไปเรียนรู้และสร้างแบบจำลองเพื่อทำนายหมวดหมู่ของสถานที่ท่องเที่ยวซึ่งงานวิจัยนี้ต้องการวัดและเปรียบเทียบประสิทธิภาพของแบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยว ซึ่งสร้างขึ้นจากข้อมูลของคุณลักษณะที่แตกต่างกันจากการสอบถามข้อมูลทั้ง 2 วิธีนี้



รูปที่ 5 กระบวนการต่าง ๆ ในขั้นตอนการเตรียมข้อมูล

2.2 การสกัดคุณลักษณะ (Feature Extraction) เป็นขั้นตอนการสกัดและเลือกเฉพาะข้อมูลที่สนใจ คือข้อมูลเกี่ยวข้องกับการจัดหมวดหมู่สถานที่ท่องเที่ยว จากข้อมูลจำนวนมหาศาลที่อยู่ในเอกสารซึ่งประกอบด้วยขั้นตอนต่าง ๆ ดังนี้

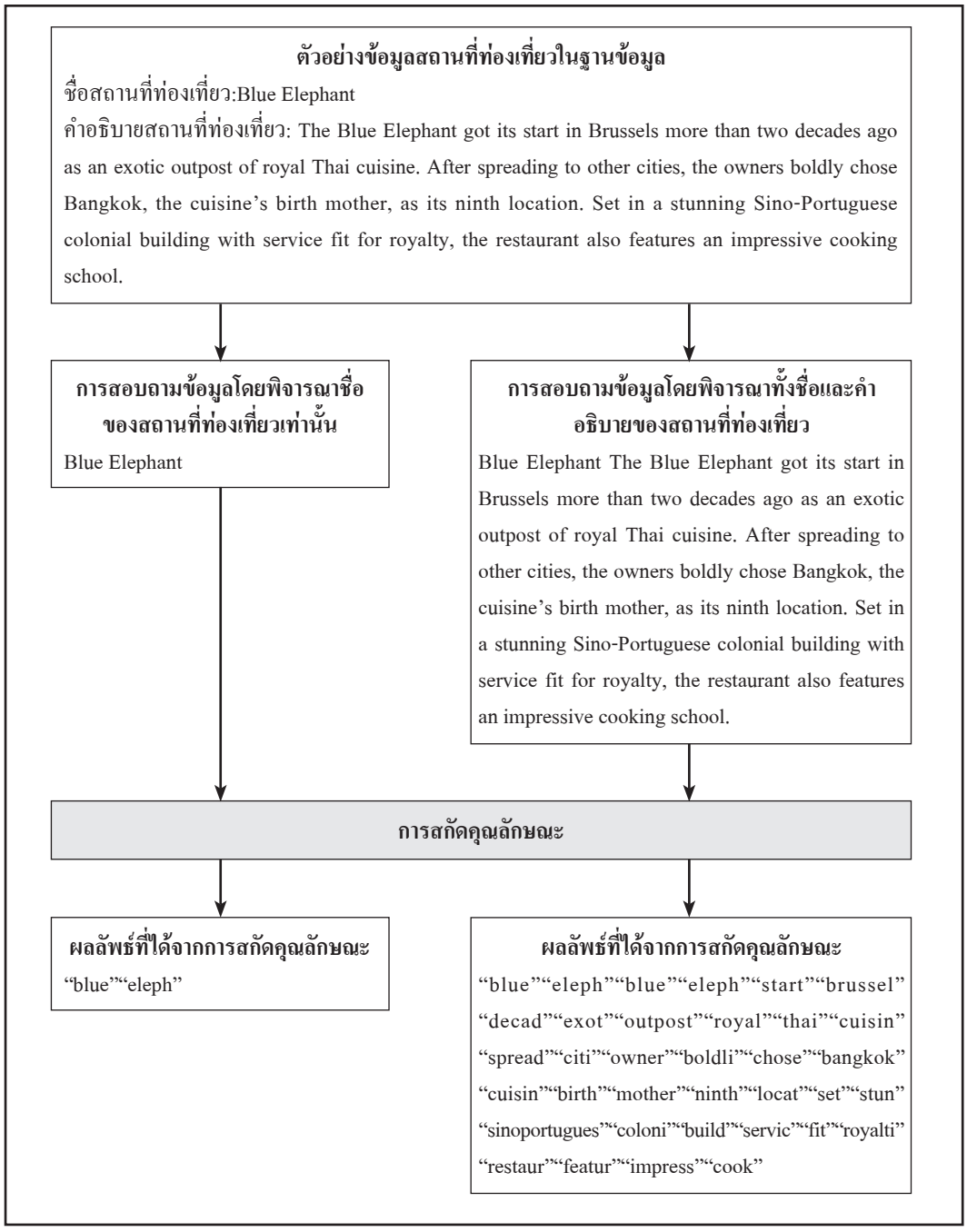
ขั้นตอนที่ 1 การสกัดคำจากข้อความ (Text Term Extraction) เป็นการแปลงข้อมูลของสถานที่ท่องเที่ยวในฐานข้อมูลให้เป็นข้อความทั่วไป (Plain Text) คือ ไม่เป็นตัวหนาตัวเอียงตัวอักษรถูกทำให้เป็นตัวเล็ก และกำจัดอักขระพิเศษและข้อความที่เกี่ยวกับเว็บไซต์เช่น แท็กเอชทีเอ็มแอล (HTML Tag)

ขั้นตอนที่ 2 การตัดคำ (Word Segmentation) เป็นการนำข้อความทั่วไปซึ่งอยู่ในรูปแบบประโยคมาแบ่งออกเป็นคำหรือคุณลักษณะ (Term/Feature) เช่น ประโยค “I like natural attractions” ถูกตัดเป็น 4 คำ คือ “I” “like” “natural” และ “attractions” ซึ่งในประโยคภาษาอังกฤษใช้เว้นวรรคในการตัดคำ

ขั้นตอนที่ 3 การกำจัดคำหยุด (Stop Word Removal) เป็นการกำจัดคำที่ไม่มีนัยสำคัญออกไป โดยไม่ทำให้ความหมายในเอกสารหรือข้อความนั้นเสียหรือเปลี่ยนแปลงไป โดยเฉพาะคำนำหน้านาม (Article) คำบุพบท (Preposition) คำสรรพนาม (Pronoun) และคำสันธาน (Conjunction) เช่น “a” “an” “the” “he” “she” “it” “and” “who” “which” และ “that” เป็นต้น

ขั้นตอนที่ 4 การหารากศัพท์ของคำ (Stemming Word) เป็นการหารูปแบบดั้งเดิมหรือรากศัพท์ของคำนั้น ๆ โดยปราศจากอุปสรรค (Prefixes) และปัจจัย (Suffixes) โดยส่วนใหญ่จะเป็นคำนามและคำกริยาเช่น คำว่า “walked” และ “walking” รูปแบบเดิมของคำ คือ “walk” แม้ว่า “ed” และ “ing” จะถูกตัดออกไปแต่ความหมายของคำ ๆ นั้นยังคงเดิม การหารากศัพท์จะช่วยรวมคำดังกล่าวให้เป็นคำเดียวกันเพื่อลดความซ้ำซ้อนของคำหรือคุณลักษณะที่เกิดขึ้นในเอกสารซึ่งงานวิจัยนี้ได้ประยุกต์ใช้ขั้นตอนวิธีการหารากศัพท์ของพอร์เตอร์ (Porter, 1980) ซึ่งเหมาะสำหรับการหารากศัพท์ของคำที่เป็นภาษาอังกฤษ

สำหรับตัวอย่างการสกัดคุณลักษณะจากการสอบถามข้อมูลสถานที่ท่องเที่ยวทั้ง 2 กรณี คือ การสอบถามเพียงแค่ชื่อของสถานที่ และการสอบถามข้อมูลทั้งชื่อและคำอธิบายของสถานที่ สามารถแสดงได้ดังตัวอย่างในรูปที่ 6



รูปที่ 6 ตัวอย่างการสกัดคุณลักษณะจากข้อมูลสถานที่ท่องเที่ยวในฐานข้อมูล

2.3 การเลือกคุณลักษณะ (Feature Selection) แม้ว่าขั้นตอนการสกัดคุณลักษณะจะมีการลดจำนวนคุณลักษณะที่ไม่สำคัญไปด้วย แต่คุณลักษณะที่ถูกสกัดออกมานั้นยังมีจำนวนมากอยู่ การนำคุณลักษณะจำนวนมากมาใช้ในการประมวลผลจะใช้เวลาและทรัพยากรในการประมวลผลมาก ดังนั้นจึงจำเป็นต้องมีการเลือกคุณลักษณะเพื่อให้ข้อมูลมีขนาดลดลง แต่ต้องสูญเสียลักษณะสำคัญของข้อมูลและความถูกต้องของผลลัพธ์ให้น้อยที่สุด สำหรับเทคนิคที่นำมาใช้ในการเลือกคุณลักษณะมีหลายเทคนิค ในงานวิจัยนี้ประยุกต์เทคนิคการเลือกคุณลักษณะด้วยวิธีการทางสถิติมาใช้ โดยใช้วิธีการให้น้ำหนักของคำในเอกสารแบบ TF/IDF (Term Frequency/Inverse Document Frequency) ซึ่งเป็นวิธีที่ใช้ประเมินความสำคัญของคำหรือคุณลักษณะนั้น ๆ ในเอกสารที่เป็นชุดฝึก (Training Set) ทั้งหมด (Jing, Huang and Shi, 2002) คุณลักษณะที่มีค่าน้ำหนัก TF/IDF มาก หมายความว่าคุณลักษณะนั้นมีประสิทธิภาพในการจำแนกเอกสารมากเช่นกัน การคำนวณค่าน้ำหนัก TF-IDF แสดงในสมการที่ 1 และ 2

$$W_{(f,d)} = TF_{(f,d)} \times IDF_{(f)} \tag{1}$$

$$IDF_{(f)} = log \frac{|D|}{|DF_{(f)}|} \tag{2}$$

โดยที่ $W_{(f,d)}$ คือค่าน้ำหนักของคุณลักษณะ (f) ที่ปรากฏในเอกสาร (d)
$TF_{(f,d)}$ คือความถี่ของคุณลักษณะ (f) ที่ปรากฏในเอกสาร (d)
$|D|$ คือจำนวนเอกสารทั้งหมดในข้อมูลชุดฝึก (Training Set)
$|DF_{(f)}|$ คือจำนวนของเอกสารทั้งหมดที่มีคุณลักษณะ (f) ปรากฏอยู่

การพิจารณาเลือกคุณลักษณะจะกระทำภายหลังจากคำนวณน้ำหนัก TF-IDF เสร็จโดยพิจารณาจากคุณลักษณะที่ให้ค่า TF-IDF สูงในแต่ละหมวดหมู่ของสถานที่ท่องเที่ยว โดยนำมารวมกันและขมวดคำที่ซ้ำกันเข้าด้วยกัน ซึ่งจำนวนคุณลักษณะที่เลือกใช้มีอยู่หลายช่วง เริ่มตั้งแต่จำนวนน้อยคือหลักสิบจนไปถึงหลักร้อย จากนั้นนำคุณลักษณะ (Feature) ที่เลือกมาแปลงให้อยู่ในรูปแบบของเมตริกซ์เอกสาร (Document-Terms Matrix) ดังตารางที่ 2 ก่อนนำไปสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่อง

ตารางที่ 2 เมตริกซ์เอกสาร

Document	Features				Class
	$Feature_1$	$Feature_2$	…	$Feature_j$	
Doc_1	F_{11}	F_{12}	…	F_{1j}	hotel
Doc_2	F_{21}	F_{22}	…	F_{2j}	restaurant
Doc_3	F_{31}	F_{32}	…	F_{3j}	nightlife
…	…	…	…	…	…
Doc_1	F_{i1}	F_{i2}	…	F_{ij}	sight seeing

2.4 การสร้างแบบจำลองการจัดหมวดหมู่ (Categorization Model Construction) การสร้างแบบจำลองเพื่อจัดการกับข้อมูลให้อยู่ในหมวดหมู่ที่ผู้ใช้กำหนด เป็นหน้าที่หลักของการทำเหมืองข้อมูลด้วยเทคนิคการจัดหมวดหมู่ (Classification) ภายในแบบจำลองประกอบไปด้วยกฎเพื่อช่วยในการตัดสินใจซึ่งสร้างมาจากข้อมูลที่มีอยู่ เพื่อใช้ในการทำนายหรือคาดการณ์ข้อมูลใหม่ในอนาคตในงานวิจัยนี้ใช้เทคนิคการเรียนรู้ของเครื่อง4เทคนิคได้แก่ต้นไม้การตัดสินใจ (Decision Trees) นาอีฟเบย์ส (Naïve Bayes) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine:SVM) และเคเนียเรสเนเบอร์ (k-Nearest Neighbor: kNN)ในการสร้างกฎการจัดหมวดหมู่สถานที่ท่องเที่ยว โดยมีการเปรียบเทียบประสิทธิภาพของแบบจำลองที่ได้จากแต่ละเทคนิค ซึ่งรายละเอียดของเทคนิคทั้ง 4 ดังต่อไปนี้

2.4.1 ต้นไม้การตัดสินใจ (Decision Tree -J48) ต้นไม้การตัดสินใจเป็นขั้นตอนวิธีหนึ่งที่ได้รับความนิยมในการจำแนกหมวดหมู่เอกสาร กฎการจำแนกหมวดหมู่ข้อมูลที่ได้จากขั้นตอนวิธีนี้อยู่ในรูปแบบของลักษณะโครงสร้างต้นไม้ ซึ่งภายในประกอบไปด้วยโหนด (Node) และกิ่ง (Branch) แต่ละโหนดจะถูกแทนด้วยคุณลักษณะ (Feature) ของชุดข้อมูลที่นำมาเรียนรู้และทดสอบ แต่ละกิ่งของต้นไม้แสดงผลในการในการทดสอบและลีฟโหนด (Leaf Node) แสดงหมวดหมู่ที่ผู้ใช้กำหนด ส่วนเกณฑ์การเลือกคุณลักษณะเพื่อนำมาเป็นโหนดของต้นไม้นั้นมาจากการคำนวณค่าเกนสารสนเทศ (Information Gain) โดยพิจารณาคุณลักษณะที่มีค่าเกนสารสนเทศหรือมีค่าเอ็นโทรพี (Entropy) ต่ำหมายความว่าคุณลักษณะนั้นมีความสามารถในการจำแนกหมวดหมู่สูง (Lee and Lee, 2006)

2.4.2 นาอีฟเบย์ส (Naïve Bayes) เป็นขั้นตอนวิธีที่ได้รับความนิยมและถูกนำมาใช้อย่างแพร่หลายในงานจำแนกหมวดหมู่เอกสารเนื่องจากความเรียบง่ายของขั้นตอนวิธีและให้ประสิทธิภาพการจำแนกที่ดีนาอีฟเบย์สเป็นขั้นตอนวิธีที่มีพื้นฐานมาจากทฤษฎีเบย์ส (Bayes' Theorem) ซึ่งอาศัยหลักความน่าจะเป็นในการทำนายผลลัพธ์ โดยการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์สำหรับรูปแบบการคำนวณความน่าจะเป็นของนาอีฟเบย์สสามารถคำนวณได้จากสมการที่ 3

$$p(Class_j | Place_i) = \frac{p(Class_j \times p(Place_i | Class_j)}{p(Place_i)} \quad (3)$$

โดยที่ $p(Class_j | Place_i)$ ความน่าจะเป็นที่สถานที่ *(Place)* ที่ i จะอยู่ในหมวดหมู่ *(Class)* ที่ j เมื่อ $1 \leq i \leq n, 1 \leq j \leq 5$ และ n คือจำนวนสถานที่ทั้งหมด

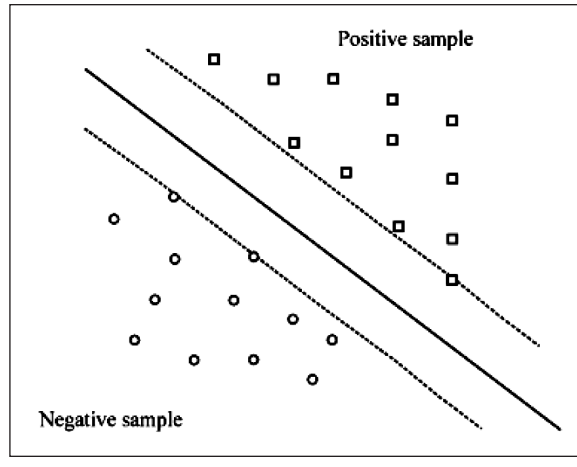
$p(Class_j)$ ความน่าจะเป็นของหมวดหมู่ *(Class)* ที่ j

$p(Place_j)$ ความน่าจะเป็นของสถานที่ *(Class)* ที่ i

$p(Place_i | Class_j)$ ความน่าจะเป็นที่คุณลักษณะ *(Class)* ของสถานที่ *(Class)* ที่ i ปรากฏในหมวดหมู่ *(Class)* ที่ j สามารถคำนวณได้จากสมการที่ 4

$$p(Place_i | Class_j) = p(f_1, f_2, ...f_n | Class_j) = \prod_k^n = 1(f_k | Class_j) \quad (4)$$

2.4.3 ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine:SVM) เป็นวิธีการจำแนกกลุ่มข้อมูลที่อาศัยระนาบการตัดสินใจมาใช้ในการแบ่งข้อมูลออกเป็น 2 ส่วน โดยพยายามสร้างเส้นแบ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตทั้งสองกลุ่มมากที่สุด ดังตัวอย่างในรูปที่ 7 ซึ่งซัพพอร์ทเวกเตอร์แมชชีนจะใช้ฟังก์ชั่นแม็พพิ้ง (Mapping Function) เพื่อแปลงข้อมูลจากโดเมนเดิมไปยังโดเมนที่เรียกว่าฟีเจอร์ สเปซ (Feature Space) และใช้ฟังก์ชั่นเคอร์เนล (Kernel Function)ในการวัดความคล้ายกันของข้อมูลในฟีเจอร์ สเปซ ในงานวิจัยนี้ใช้เคอร์เนลฟังก์ชั่น คือ โพลิโนเมียล เคอร์เนล (Polynomial Kernel) ในการทดสอบข้อมูลด้วยเทคนิคนี้



รูปที่ 7 ตัวอย่างระนาบการตัดสินใจของซัพพอร์ทเวกเตอร์แมชชีน
(Ali, Shamsuddin and Ismail, 2011)

2.4.4 เคเนียเรสเนเบอร์ (k-Nearest Neighbor:kNN) เป็นเทคนิคที่ใช้ในการจัดหมวดหมู่ข้อมูลที่เรียบง่ายและมีประสิทธิภาพ แต่เทคนิคนี้มีวิธีการจัดหมวดหมู่ที่แตกต่างจากเทคนิคอื่นตรงที่ไม่ได้ใช้ข้อมูลชุดฝึกในการสร้างแบบจำลอง แต่จะใช้ข้อมูลทั้งชุดมาเป็นตัวแบบจำลองเลย หลักการจัดหมวดหมู่ข้อมูลจะพิจารณาจากค่า kเช่น ค่า $k = 5$ หมายถึง ขั้นตอนวิธีจะพิจารณาข้อมูลจำนวน 5 ชุด มีลักษณะใกล้เคียงกับข้อมูลชุดตัวอย่าง การคำนวณความใกล้เคียงจะใช้วิธีการวัดระยะทางยูคลิเดียน (Euclidian Distance) จากสมการที่ 5 ภายหลังจากที่รวบรวมจำนวนสมาชิกที่ใกล้เคียงที่สุด k ตัวได้แล้ว โดยการเลือกหมวดหมู่จะเลือกหมวดหมู่ที่สมาชิกส่วนใหญ่ที่อยู่ในกลุ่ม k สังกัดอยู่มากที่สุดให้กับสมาชิกใหม่ที่มาเข้ากลุ่มด้วย ในงานวิจัยนี้กำหนดค่า k เท่ากับ 1

$$dist\,(X_i, X_j) = \sqrt{\sum_{k=1}^{n} (X_{i,K} - X_{j,K})^2} \tag{5}$$

โดยที่ $dist\,(X_i, X_j)$ คือ ระยะห่างระหว่างข้อมูลตัวอย่าง X_i กับข้อมูลตัวอย่าง X_j
n คือ จำนวนคุณลักษณะทั้งหมดของข้อมูลตัวอย่าง
$X_{i,k}$ คือ คุณลักษณะตัวที่ k ของข้อมูลตัวอย่าง X_i

2.5 การประเมินแบบจำลอง (Model Evaluation) เป็นการวัดประสิทธิภาพในการทำนายหมวดหมู่สถานที่ท่องเที่ยวของแบบจำลอง ด้วยเกณฑ์ทางด้านการค้นคืนสารสนเทศ คือ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) นอกจากนี้ยังประเมินในส่วนของเวลาที่ใช้ในการสร้างแบบจำลองของแต่ละเทคนิคด้วย สำหรับการคำนวณค่าความถูกต้อง ค่าความแม่นยำและค่าความระลึกสามารถแสดงได้ดังสมการที่ 6, 7 และ 8

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

โดยที่ TP = สถานที่ท่องเที่ยวที่อยู่ในหมวดหมู่ C_j และแบบจำลองทำนายว่าอยู่ในหมวดหมู่ C_j
FP = สถานที่ท่องเที่ยวที่ไม่อยู่ในหมวดหมู่ C_j และแบบจำลองทำนายว่าอยู่ในหมวดหมู่ C_j
FN = สถานที่ท่องเที่ยวที่อยู่ในหมวดหมู่ C_j และแบบจำลองทำนายว่าไม่อยู่ในหมวดหมู่ C_j

TN = สถานที่ท่องเที่ยวที่ไม่อยู่ในหมวดหมู่ C_j และแบบจำลองทำนายว่าไม่อยู่ในหมวดหมู่C_j

Cj = กลุ่มหมวดหมู่ของสถานที่ท่องเที่ยว 5 หมวดหมู่ เมื่อ $1 \leq j \leq 5$

การทดสอบแบบจำลอง

1. สภาพแวดล้อมในการทดสอบ

ในการทดสอบแบบจำลองนั้น ใช้เครื่องคอมพิวเตอร์ซึ่งเป็นระบบปฏิบัติการไมโครซอฟต์วินโดวส์เซเว่น รุ่นโปรเฟสชั่นนอล (Microsoft Windows 7 OS Professional) หน่วยประมวลผลอินเทล คอร์ไอไฟว์ (Intel Core i5) ความเร็ว 2.30 กิกะเฮิรตซ์ (GHz) หน่วยความจำ 4 กิกะไบต์ (GB) และพื้นที่จัดเก็บข้อมูล 500 กิกะไบต์ พร้อมการเชื่อมต่ออินเทอร์เน็ตแบบมีสายและไร้สาย ส่วนภาษาการเขียนโปรแกรม (Programming Language) ที่ใช้ในการประมวลผลภาษาธรรมชาติใช้ภาษาพีเอชพี (PHP) รุ่น 5.2.6 ร่วมกับฐานข้อมูลมายเอสคิวแอล (MySQL) รุ่น 5.0.51b สำหรับซอฟต์แวร์ที่ใช้ในการประมวลผล ทดสอบ และประเมินผลแบบจำลองที่ได้จากการเรียนรู้ของเครื่อง คือ เวก้าสำหรับระบบปฏิบัติการวินโดวส์ (Weka for Windows OS) รุ่น 3.65

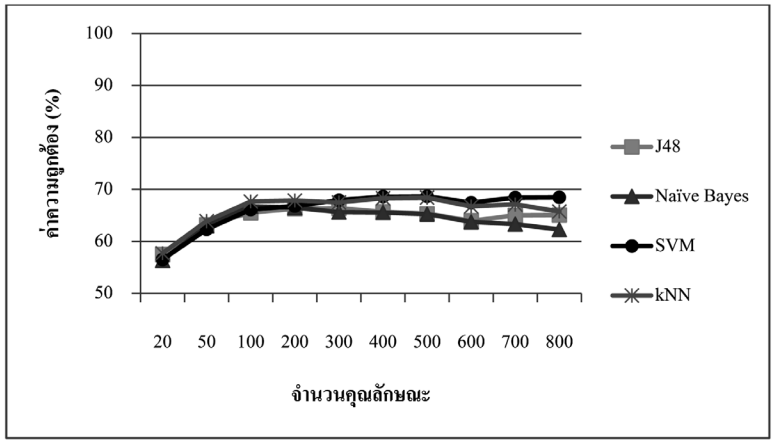
ข้อมูลที่นำมาใช้เป็นชุดฝึกและชุดทดสอบในการสร้างแบบจำลอง คือ ข้อมูลสถานที่ท่องเที่ยวในประเทศไทยที่รวบรวมมาจากเว็บไซต์โลนลี่แพลนเน็ต มีจำนวนหมวดหมู่ของสถานที่ท่องเที่ยวทั้งสิ้น 5 หมวดหมู่ และจำนวนระเบียนข้อมูลรวมทั้งสิ้น 1,606 ระเบียน และนำมาแปลงให้อยู่ในรูปแบบเมตริกซ์เอกสารก่อนนำไปสร้างแบบจำลองสำหรับการสร้างแบบจำลองมีการแบ่งข้อมูลชุดฝึกและชุดทดสอบโดยใช้วิธีการตรวจสอบแบบไขว้กัน k ชุด (*k*-Fold Cross Validation) ซึ่งข้อมูลจะถูกแบ่งออกเป็น *k* ชุดจำนวนเท่า ๆ กัน และทำการคำนวณตรวจสอบจำนวน *k* รอบ โดยแต่ละรอบข้อมูลชุดที่ *k* จะถูกเลือกออกมาให้เป็นข้อมูลชุดทดสอบ ส่วนข้อมูลชุดที่เหลือซึ่งมีจำนวนเท่ากับ *k*-1ชุด จะถูกใช้เป็นข้อมูลชุดฝึก (Kohavi, 1995) ซึ่งในการทดสอบนี้ใช้วิธีการตรวจสอบแบบไขว้กัน 10ชุด (10-Fold Cross Validation)

2. ผลการทดสอบและการอภิปรายผล

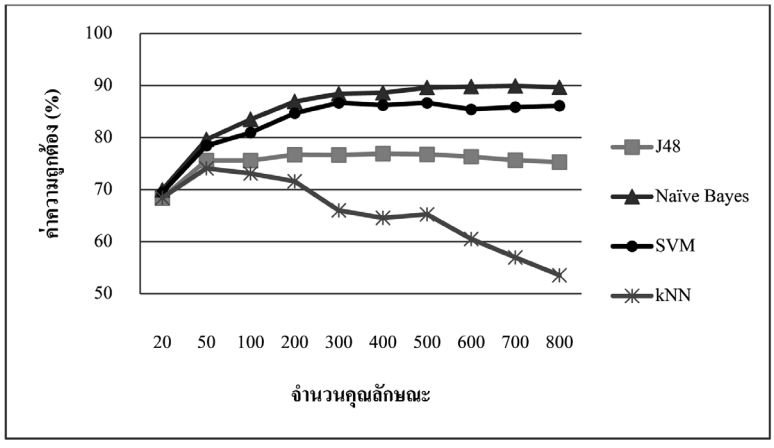
จากการสร้างแบบจำลองเพื่อจัดหมวดหมู่สถานที่ท่องเที่ยวโดยใช้เทคนิคการเรียนรู้ของเครื่องสามารถทดสอบและอภิปรายผลโดยแบ่งเป็น 4 ประเด็นหลัก ได้แก่ การทดสอบความถูกต้องในการจัดหมวดหมู่ การทดสอบระยะเวลาที่ใช้ในการสร้างแบบจำลอง การทดสอบประสิทธิภาพด้านการค้นคืนสารสนเทศและการทดสอบความถูกต้องในการจัดหมวดหมู่กับข้อมูลชุดอื่นซึ่งมีรายละเอียดดังต่อไปนี้

ประเด็นที่ 1 การทดสอบความถูกต้องในการจัดหมวดหมู่ประเด็นแรกเป็นการทดสอบค่าความถูกต้องของแบบจำลองต่าง ๆ ซึ่งใช้เทคนิคการเรียนรู้ของเครื่องที่แตกต่างกัน 4 เทคนิค และการเลือกคุณลักษณะของข้อมูลที่นำมาสร้างแบบจำลองที่แตกต่างกัน 2 กรณี คือ1) การใช้ชื่อของสถานที่

ท่องเที่ยวเท่านั้น และ 2) การใช้ทั้งชื่อและคำอธิบายของสถานที่ท่องเที่ยว โดยใช้เกณฑ์ค่าความถูกต้อง (Accuracy) ในการประเมินผลในรูปที่8และ 9แสดงผลการทดสอบดังกล่าว ซึ่งถ้าเปรียบเทียบกันแล้วพบว่าการสร้างแบบจำลองโดยพิจารณาชื่อและคำอธิบายของสถานที่ท่องเที่ยว (ในรูปที่ 9) จะให้ค่าความถูกต้องที่สูงกว่าการใช้ชื่อสถานที่ท่องเที่ยวเพียงอย่างเดียว (ในรูปที่ 8)



รูปที่ 8 ผลการทดสอบความถูกต้องในการจัดหมวดหมู่สถานที่ท่องเที่ยวของแบบจำลองที่สร้างจากคุณลักษณะของชื่อสถานที่ท่องเที่ยวในเทคนิคการเรียนรู้ของเครื่อง 4 เทคนิค



รูปที่ 9 ผลการทดสอบความถูกต้องในการจัดหมวดหมู่สถานที่ท่องเที่ยวของแบบจำลองที่สร้างจากคุณลักษณะของชื่อและคำอธิบายของสถานที่ท่องเที่ยวในเทคนิคการเรียนรู้ของเครื่อง 4 เทคนิค

ค่าความถูกต้องที่ปรากฏในรูปที่8ซึ่งเป็นการเลือกคุณลักษณะจากชื่อสถานที่ท่องเที่ยวเท่านั้น จะเห็นได้ว่าไม่มีเทคนิคใดที่สามารถให้ค่าความถูกต้องได้ถึง 70% ทุกเทคนิคจะให้ค่าความถูกต้องโดยเฉลี่ยเท่า ๆ กันที่จำนวนคุณลักษณะเดียวกันสำหรับค่าความถูกต้องที่ปรากฏในรูปที่ 9 เป็นการใช้ทั้งชื่อและคำอธิบายสถานที่ท่องเที่ยวมาพิจารณาสร้างแบบจำลอง โดยผลการทดสอบปรากฏว่าเกือบทุกเทคนิคให้ประสิทธิภาพการจัดหมวดหมู่ที่ดีกว่าการเลือกคุณลักษณะในกรณีแรก คือ มากกว่า 70% ตั้งแต่จำนวนคุณลักษณะ 50 คุณลักษณะ ยกเว้นเทคนิคเคเนียเรสเนเบอร์ (kNN) ที่เริ่มมีประสิทธิภาพลดลงที่ตั้งแต่จำนวนมากกว่า 50 คุณลักษณะเป็นต้นไปสำหรับต้นไม้การตัดสินใจ (J48) มีแนวโน้มที่จะให้ค่าความถูกต้องลดลงเมื่อมีจำนวนคุณลักษณะมากกว่า 400 คุณลักษณะ ซัพพอร์ทเวกเตอร์แมชชีนให้ค่าความถูกต้องที่เพิ่มขึ้นเรื่อย ๆ ตามจำนวนคุณลักษณะที่เพิ่มมากขึ้นจนกระทั่งถึง 600 คุณลักษณะ ค่าความถูกต้องที่ได้นั้นลดลง แต่หลังจากจำนวนมากกว่า 600 คุณลักษณะขึ้นไปกลับมีแนวโน้มเพิ่มขึ้น ส่วนนาอีฟเบย์สนั้นให้ค่าความถูกที่เพิ่มขึ้นเรื่อยๆ เมื่อจำนวนคุณลักษณะเพิ่มมากขึ้นและเริ่มคงที่ที่ 89% ณ จำนวน 500 คุณลักษณะขึ้นไป จากกราฟจะเห็นได้ว่าเทคนิคนาอีฟเบย์สและซัพพอร์ทเวกเตอร์แมชชีนยิ่งมีจำนวนคุณลักษณะมากเท่าไหร่ค่าความถูกต้องที่ได้ยิ่งสูง ในกรณีนี้อาจเกิดปัญหาการจัดหมวดหมู่ที่คุณลักษณะที่ใช้การจำแนกหมวดหมู่พอดีเกินไปกับข้อมูลชุดฝึกและชุดทดสอบ (Overfitting) แต่อาจไม่เหมาะกับข้อมูลที่ต้องการทำนายจริง ดังนั้นการเลือกจำนวนคุณลักษณะที่จะนำมาสร้างแบบจำลองด้วยเทคนิคนาอีฟเบย์สและซัพพอร์ทเวกเตอร์แมชชีนถ้าผู้ใช้ต้องการความถูกต้องที่ 80% ขึ้นไปควรพิจารณาใช้จำนวนคุณลักษณะที่100-500คุณลักษณะเท่านั้นและหากเรียงตามลำดับค่าความถูกต้องแล้วเทคนิคนาอีฟเบย์สจะให้ค่าความถูกต้องสูงที่สุด รองลงมาคือซัพพอร์ทเวกเตอร์แมชชีน ต้นไม้การตัดสินใจและเคเนียเรสเนเบอร์ตามลำดับ

โดยสรุปแล้ว การเลือกคุณลักษณะโดยพิจารณาทั้งชื่อและคำอธิบายของสถานที่ท่องเที่ยวมาสร้างแบบจำลองจะเห็นได้ชัดเจนว่าค่าความถูกต้องในการจัดหมวดหมู่สถานที่ท่องเที่ยวที่ได้จากแบบจำลองทุกเทคนิคสูงกว่าการเลือกคุณลักษณะโดยพิจารณาเพียงชื่อสถานที่ท่องเที่ยวอย่างเดียวมาสร้างแบบจำลอง ซึ่งจำนวนของคุณลักษณะต่าง ๆ นั้นมีผลต่อค่าความถูกต้องในแต่ละเทคนิค ดังนั้นการเลือกจำนวนคุณลักษณะเพื่อนำมาใช้สร้างแบบจำลองควรพิจารณาจำนวนคุณลักษณะที่เหมาะสมหรือให้ค่าความถูกต้องที่ตรงตามความพอใจของผู้ใช้และไม่ก่อให้เกิดปัญหาการจัดหมวดหมู่ดังกล่าว

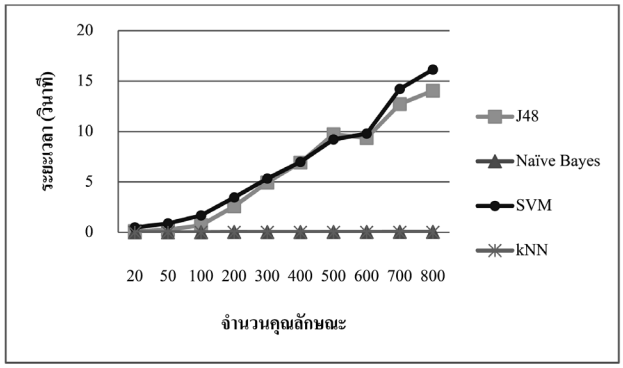
ประเด็นที่ 2 การทดสอบระยะเวลาที่ใช้ในการสร้างแบบจำลอง ประเด็นนี้เป็นการทดสอบเกี่ยวกับระยะเวลาในการสร้างแบบจำลองของเทคนิคการเรียนรู้ของเครื่องทั้ง 4 เพื่อดูว่าในกรณีที่จำนวนคุณลักษณะมากน้อยต่างกันนั้นมีผลต่อระยะเวลาในการสร้างแบบจำลองมากน้อยเพียงใด รูปที่ 10 และ 11 แสดงผลการสร้างแบบจำลองของแต่ละเทคนิค โดยแบ่งการทดสอบออกเป็นสองกรณีเช่นเดียวกันประเด็นที่ 1 คือ การใช้เพียงชื่อสถานที่ท่องเที่ยวดังแสดงในรูปที่10และการใช้ชื่อกับคำอธิบายของสถานที่ท่องเที่ยวดังแสดงในรูปที่ 11

จากรูปที่ 10 และ 11 ผลการทดสอบที่ทั้งสองกรณีมีร่วมกันในการทดสอบระยะเวลาที่ใช้ในการสร้างแบบจำลองมี 2 แง่มุม คือ

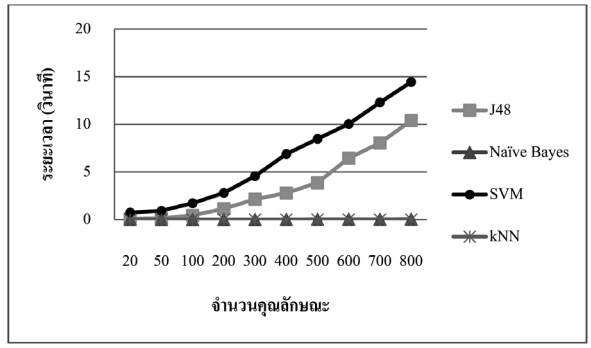
แง่มุมที่ 1 เทคนิคเคเนียเรสเนเบอร์และนาอีฟเบย์ส จำนวนคุณลักษณะมากหรือน้อยไม่ส่งผลต่อเวลาที่ใช้ในการสร้างแบบจำลองของเทคนิคทั้งสอง โดยนาอีฟเบย์สใช้เวลาสร้างแบบจำลองในทุกจำนวนคุณลักษณะที่ทดสอบในการเลือกคุณลักษณะกรณีแรกเฉลี่ย 0.031 วินาที และในการเลือกคุณลักษณะกรณีที่สองเฉลี่ย 0.021 วินาที สำหรับเคเนียเรสเนเบอร์ เวลาที่ใช้ในการสร้างแบบจำลองจะเป็น 0 วินาที ทั้งหมด เนื่องจากเทคนิคนี้ไม่ได้มีการนำข้อมูลชุดฝึกมาใช้สร้างแบบจำลอง

แง่มุมที่ 2 เทคนิคซัพพอร์ทเวกเตอร์แมชชีนและต้นไม้การตัดสินใจ นั้นจำนวนคุณลักษณะยิ่งมากเวลาที่ใช้ในการสร้างแบบจำลองยิ่งนานขึ้นตามไปด้วย แต่อย่างไรก็ตาม เมื่อพิจารณาตามการเลือกคุณลักษณะที่แตกต่างกันพบว่า ระยะเวลาที่ใช้ในการสร้างแบบจำลองของเทคนิคซัพพอร์ทเวกเตอร์แมชชีนและต้นไม้การตัดสินใจ ซึ่งใช้คุณลักษณะกรณีแรกจะใช้เวลาในการสร้างแบบจำลองนานกว่าในกรณีที่สอง หากเปรียบเทียบให้ชัดเจนในกรณีแรกที่จำนวน 800 คุณลักษณะ เทคนิคซัพพอร์ทเวกเตอร์แมชชีนใช้ระยะเวลา 16.14 วินาที และต้นไม้การตัดสินใจ ใช้ระยะเวลา 14.05 วินาที ในขณะที่กรณีที่สองซัพพอร์ทเวกเตอร์แมชชีนใช้ระยะเวลาเพียง14.45วินาที และต้นไม้การตัดสินใจ ใช้ระยะเวลาเพียง 10.4 วินาที

สรุปได้ว่าในกรณีที่จำนวนคุณลักษณะเท่ากันการเลือกคุณลักษณะโดยใช้ชื่อสถานที่ท่องเที่ยวเท่านั้น ใช้ระยะเวลาสร้างแบบจำลองนานกว่าการเลือกคุณลักษณะโดยใช้ทั้งชื่อและคำอธิบายสถานที่ท่องเที่ยวในเทคนิคซัพพอร์ทเวกเตอร์แมชชีนและต้นไม้การตัดสินใจ ส่วนเทคนิคเคเนียเรสเนเบอร์และนาอีฟเบย์สนั้น วิธีการพิจารณาคุณลักษณะทั้ง 2 กรณีได้ผลลัพธ์ที่เหมือนกันทั้งคู่อีกทั้งจำนวนคุณลักษณะมากน้อยไม่มีผลต่อระยะเวลาในการสร้างแบบจำลองของเทคนิคทั้งสองมากนัก โดยเฉพาะเทคนิคเคเนียเรสเนเบอร์ที่ไม่ได้นำข้อมูลชุดฝึกมาใช้สร้างแบบจำลองทำให้ไม่มีระยะเวลาที่ใช้ในการสร้างแบบจำลอง

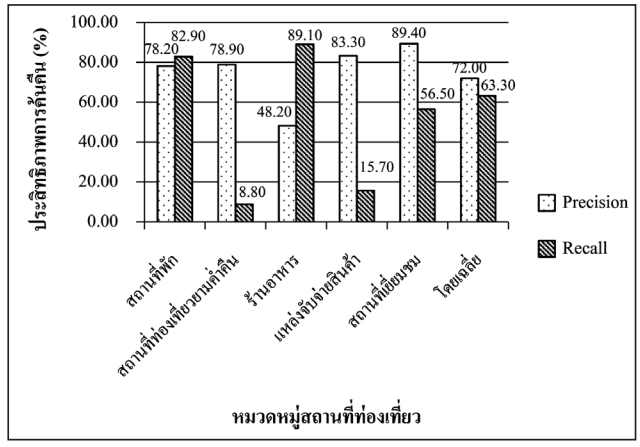


รูปที่ 10 ผลการทดสอบระยะเวลาที่ใช้ในการสร้างแบบจำลองที่สร้างจากคุณลักษณะของชื่อสถานที่ท่องเที่ยวในเทคนิคการเรียนรู้ของเครื่อง 4 เทคนิค



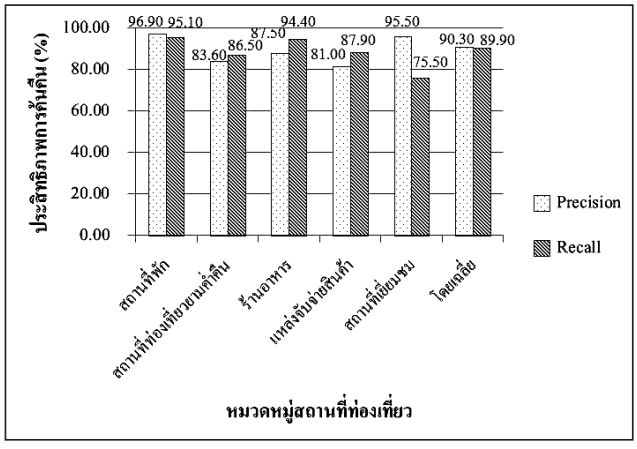
รูปที่ 11 ผลการทดสอบระยะเวลาที่ใช้ในการสร้างแบบจำลองที่สร้างจากคุณลักษณะของชื่อและคำอธิบายของสถานที่ท่องเที่ยวในเทคนิคการเรียนรู้ของเครื่อง 4 เทคนิค

ประเด็นที่ 3 การทดสอบประสิทธิภาพด้านการค้นคืนสารสนเทศจากผลการทดสอบในประเด็นที่ 1 และ 2 เทคนิคนาอีฟเบย์สได้แสดงให้เห็นถึงค่าความถูกต้องและเวลาที่ใช้ในการสร้างแบบจำลองที่ดีกว่าเทคนิคอื่น ๆ ดังนั้นในประเด็นที่ 3 นี้จึงนำเพียงเทคนิคนาอีฟเบย์สมาเป็นตัวแทนการทดสอบประสิทธิภาพในการค้นคืนสถานที่ท่องเที่ยวทั้ง 5 หมวดหมู่ โดยเปรียบเทียบจากกรณีการเลือกคุณลักษณะมาสร้างแบบจำลองทั้ง 2 กรณีที่แตกต่างกัน คือ การใช้เพียงชื่อสถานที่ท่องเที่ยว และการใช้ชื่อกับคำอธิบายของสถานที่ท่องเที่ยว จำนวนคุณลักษณะที่นำมาใช้สร้างแบบจำลองมีจำนวนที่เท่ากัน คือ 700 คุณลักษณะสำหรับการวัดประสิทธิภาพนั้นวัดจากค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ที่ได้จากเทคนิคนาอีฟเบย์ส ซึ่งผลการทดสอบแบ่งออกเป็น 2 กรณี ตามวิธีการเลือกคุณลักษณะที่นำมาสร้างแบบจำลองดังกล่าวโดยผลการเลือกคุณลักษณะโดยใช้เพียงชื่อของสถานที่ท่องเที่ยวแสดงในรูปที่ 12 และการใช้ชื่อและคำอธิบายของสถานที่ท่องเที่ยวแสดงในรูปที่ 13



รูปที่ 12 แผนภูมิสรุปผลประสิทธิภาพด้านการค้นคืนสารสนเทศจากแบบจำลองที่สร้างจากคุณลักษณะของชื่อสถานที่ท่องเที่ยวด้วยเทคนิคนาอีฟเบย์ส

ประสิทธิภาพด้านการค้นคืนสารสนเทศในรูปที่ 12 แสดงให้เห็นว่าในกรณีการนำข้อมูลเฉพาะชื่อสถานที่ท่องเที่ยวมาพิจารณาในการเลือกคุณลักษณะที่นำมาใช้ในการสร้างแบบจำลอง ส่งผลให้การค้นคืนสถานที่ท่องเที่ยวบางหมวดหมู่มีค่าความแม่นยำและค่าความระลึกที่แตกต่างกันมาก โดยเฉพาะหมวดหมู่สถานที่ท่องเที่ยวยามค่ำคืน ซึ่งมีค่าความแม่นยำต่อค่าความระลึกเท่ากับ 78.90% ต่อ 8.80% หมวดหมู่แหล่งจับจ่ายสินค้า เท่ากับ 83.30% ต่อ 15.70% และหมวดหมู่สถานที่เยี่ยมชม เท่ากับ 89.40% ต่อ 56.50% ส่วนหมวดหมู่ร้านอาหารและสถานที่พักนั้น ให้ค่าความแม่นยำที่น้อยกว่าค่าความระลึกคือ 48.20% ต่อ 89.10% และ 78.20% ต่อ 82.90% ตามลำดับโดยเฉลี่ยค่าความแม่นยำและค่าความระลึกในกรณีแรกอยู่ที่ 72.00% ต่อ 63.30%



รูปที่ 13 แผนภูมิสรุปผลประสิทธิภาพด้านการค้นคืนสารสนเทศจากแบบจำลองที่สร้างจากคุณลักษณะของชื่อและคำอธิบายของสถานที่ท่องเที่ยวด้วยเทคนิคนาอีฟเบย์ส

ส่วนรูปที่ 13 แสดงให้เห็นถึงประสิทธิภาพด้านการค้นคืนสารสนเทศที่เพิ่มมากขึ้นจากการนำชื่อและคำอธิบายของสถานที่ท่องเที่ยวเข้ามาพิจารณาร่วมกันในการเลือกคุณลักษณะที่นำมาใช้ในการสร้างแบบจำลอง คือทำให้ทั้งค่าความแม่นยำและค่าความระลึกในทุกหมวดหมู่เพิ่มขึ้นเมื่อเปรียบเทียบกับกรณีแรกโดยเฉพาะหมวดหมู่สถานที่ท่องเที่ยวยามค่ำคืนมีค่าความแม่นยำต่อค่าความระลึกเท่ากับ 83.60% ต่อ 86.50% หมวดหมู่แหล่งจับจ่ายสินค้าเท่ากับ 81.80% ต่อ 87.90% หมวดหมู่สถานที่เยี่ยมชม เท่ากับ 95.50% ต่อ 75.50% หมวดหมู่สถานที่พักเท่ากับ 96.90% ต่อ 95.10% และหมวดหมู่ร้านอาหารเท่ากับ 87.50% ต่อ 94.40% จากทุกหมวดหมู่มีเพียงแหล่งจับจ่ายสินค้าเท่านั้นที่แม้ว่าค่าความระลึกจะเพิ่มขึ้นแต่ค่าความแม่นยำกลับลดลงเพียงเล็กน้อย

เมื่อเทียบกับค่าความแม่นยำที่ได้จากกรณีแรกในหมวดหมู่เดียวกันโดยเฉลี่ยแล้วค่าความแม่นยำและค่าความระลึกที่ได้จากการเลือกคุณลักษณะมาใช้สร้างแบบจำลองในกรณีที่สองอยู่ที่ 90.30% ต่อ 89.90%

จากผลการทดสอบดังกล่าวชี้ให้เห็นว่า การนำคำอธิบายสถานที่ท่องเที่ยวเข้ามาพิจารณาในการเลือกคุณลักษณะร่วมกับชื่อนั้นส่งผลให้ประสิทธิภาพของแบบจำลองในการค้นคืนหมวดหมู่ของสถานที่ท่องเที่ยวโดยเฉลี่ยทุกหมวดหมู่มีค่าความแม่นยำและค่าความระลึกเพิ่มมากขึ้นเมื่อเทียบกับการเลือกคุณลักษณะในกรณีแรกที่เป็นการพิจารณาจากชื่อสถานที่ท่องเที่ยวเท่านั้น ดังนั้นการที่ค่าความแม่นยำและค่าความระลึกที่เพิ่มขึ้นนั่นหมายถึง แบบจำลองที่ได้มีประสิทธิภาพในการทำนายข้อมูลจริงที่เข้ามาได้อย่างถูกต้องมากยิ่งขึ้น

ประเด็นที่ 4 การทดสอบความถูกต้องในการจัดหมวดหมู่กับข้อมูลชุดอื่น ประเด็นที่ 4 นี้เป็นการทดสอบค่าความถูกต้องของแบบจำลองต่าง ๆ ซึ่งใช้เทคนิคการเรียนรู้ของเครื่องที่แตกต่างกัน 4 เทคนิค กับข้อมูลชุดอื่น ซึ่งเป็นข้อมูลสถานที่ท่องเที่ยวภาษาอังกฤษในประเทศสิงคโปร์จำนวน 639 ระเบียน จากเว็บไซต์โลนลี่ แพลนเน็ต (Lonely Planet) จำนวน 5 หมวดหมู่ ดังที่แสดงในตารางที่ 3 โดยใช้ชื่อและคำอธิบายของสถานที่ท่องเที่ยว มาสกัดคุณลักษณะจำนวน 700 คุณลักษณะในการสร้างแบบจำลอง ซึ่งคุณลักษณะจำนวน 700 คุณลักษณะนี้เป็นจำนวนที่เทคนิคนาอีฟเบย์สให้ค่าความถูกต้องสูงที่สุด วัตถุประสงค์ในประเด็นนี้เพื่อทดสอบกรณีที่ข้อมูลมีการเปลี่ยนแปลง เทคนิคการเรียนรู้ของเครื่องต่าง ๆ ยังคงให้ค่าความถูกต้องเหมือนเดิมหรือเปลี่ยนไปโดยเฉพาะเทคนิคนาอีฟเบย์สและซัพพอร์ทเวกเตอร์สำหรับผลการเปรียบเทียบค่าความถูกต้องที่ได้จากแบบจำลองจากชุดข้อมูลสถานที่ท่องเที่ยวในประเทศไทยและชุดข้อมูลสถานที่ท่องเที่ยวในประเทศสิงคโปร์แสดงในตารางที่ 4

ตารางที่ 3 จำนวนระเบียนในสถานที่ท่องเที่ยวในประเทศสิงคโปร์แต่ละหมวดหมู่

หมวดหมู่สถานที่ท่องเที่ยว	จำนวนระเบียน
สถานที่พัก (Hotel)	121
สถานที่ท่องเที่ยวยามค่ำคืน (Nightlife)	61
ร้านอาหาร (Restaurant)	212
แหล่งจับจ่ายสินค้า (Shopping)	131
สถานที่เยี่ยมชม (Sight Seeing)	114
รวม	**639**

ตารางที่ 4 ผลการเปรียบเทียบค่าความถูกต้องที่ได้จากเทคนิคการเรียนรู้ของเครื่องทั้ง 4 ระหว่างชุดข้อมูลสถานที่ท่องเที่ยวในประเทศไทยและประเทศสิงคโปร์

เทคนิคการเรียนรู้ของเครื่อง	ค่าความถูกต้อง		
	ชุดข้อมูลสถานที่ท่องเที่ยวในประเทศไทย (%)	ชุดข้อมูลสถานที่ท่องเที่ยวในประเทศสิงคโปร์ (%)	โดยเฉลี่ย (%)
ต้นไม้การตัดสินใจ	75.65	66.67	71.16
นาอีฟเบย์ส	89.91	87.48	88.70
ซัพพอร์ทเวกเตอร์แมชชีน	85.87	85.76	85.82
เคเนียเรสเนเบอร์	56.97	33.18	45.08

ผลการทดสอบเปรียบเทียบค่าความถูกต้องของแต่ละเทคนิคการเรียนรู้ของเครื่องระหว่างชุดข้อมูลสถานที่ท่องเที่ยวในประเทศไทยและประเทศสิงคโปร์ ดังปรากฏในตารางที่ 4 แสดงให้เห็นว่ากรณีที่ข้อมูลที่นำมาใช้สร้างและทดสอบแบบจำลองเป็นข้อมูลคนละชุดเทคนิคนาอีฟเบย์สยังคงเป็นเทคนิคที่ให้ค่าความถูกต้องสูงที่สุดโดยเฉลี่ยที่ 88.70% รองลงมาคือซัพพอร์ทเวกเตอร์แมชชีนต้นไม้การตัดสินใจและเคเนียเรสเนเบอร์โดยเฉลี่ยที่ 85.82% 71.16% และ 45.08% ตามลำดับ สรุปได้ว่าในกรณีที่มีการเปลี่ยนแปลงชุดข้อมูลที่นำมาใช้สร้างและทดสอบแบบจำลอง กรณีนี้ไม่ได้ส่งผลให้ประสิทธิภาพในการเรียนรู้และทำนายหมวดหมู่ของสถานที่ท่องเที่ยวของเทคนิคที่ให้ผลดีอย่างเทคนิคนาอีฟเบย์สและซัพพอร์ทเวกเตอร์แมชชีนลดลงไปแต่อย่างใด เทคนิคทั้งสองยังคงสามารถให้ค่าความถูกต้องที่สูงอยู่ได้

บทสรุป

งานวิจัยได้นำเสนอการพัฒนาแบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยวโดยใช้การเรียนรู้ของเครื่อง โดยเริ่มตั้งแต่การพัฒนาโมดูลเพื่อช่วยในการรวบรวมข้อมูลสถานที่ท่องเที่ยวแบบอัตโนมัติจากแหล่งข้อมูลออนไลน์อย่างเว็บไซต์โลนลี่แพลนเนตการเปรียบเทียบระหว่างกรณีการนำเพียงชื่อสถานที่ท่องเที่ยว และการนำทั้งชื่อและคำอธิบายของสถานที่ท่องเที่ยวมาใช้ในการเลือกคุณลักษณะที่นำมาใช้ในการสร้างแบบจำลองรวมทั้งการเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้เทคนิคการเรียนรู้ของเครื่องที่แตกต่างกัน4 เทคนิค ได้แก่ ต้นไม้การตัดสินใจ นาอีฟเบย์ส ซัพพอร์ทเวกเตอร์แมชชีน และเคเนียเรสเนเบอร์ ด้วยการวัดค่าความถูกต้อง ระยะเวลาที่แต่ละขั้นตอนวิธีใช้ในการสร้างแบบจำลองและการทดสอบวัดค่าความถูกต้องกับข้อมูลชุดอื่น ส่วนค่าความแม่นยำ และค่าความระลึกจะทดสอบกับเทคนิคนาอีฟเบย์สเท่านั้นเนื่องจากเป็นเทคนิคที่ได้ค่าความถูกต้องโดยให้ค่าความถูกต้องโดยเฉลี่ยที่ 88.70% ณ จำนวน 700 คุณลักษณะโดยจำนวน

คุณลักษณะยิ่งมากค่าความถูกต้องที่ได้จากเทคนิคนาอีฟเบย์สจะมากตามไปด้วยแต่สำหรับการเลือกจำนวนคุณลักษณะมาใช้กับเทคนิคนี้ ผู้ใช้ควรเลือกจำนวนคุณลักษณะที่ให้ค่าความถูกต้องตามที่ผู้ใช้พอใจ และไม่ก่อให้เกิดปัญหาการจัดหมวดหมู่ที่คุณลักษณะที่ใช้การจำแนกหมวดหมู่พอดีเกินไปกับข้อมูลชุดฝึกและชุดทดสอบ (Overfitting) แต่อาจไม่เหมาะกับข้อมูลที่ต้องการทำนายจริงสำหรับการทดสอบวัดค่าความถูกต้องกับข้อมูลชุดอื่นชี้ให้เห็นว่ากรณีที่ข้อมูลเป็นข้อมูลคนละชุดเทคนิคนาอีฟเบย์สและซัพพอร์ทเวกเตอร์แมชชีนยังคงเป็นเทคนิคที่สามารถให้ค่าความถูกต้องที่สูงได้

นอกจากนี้ผลการทดสอบเกี่ยวกับการเลือกคุณลักษณะพบว่าการใช้ทั้งชื่อและคำอธิบายของสถานที่ท่องเที่ยวมาใช้ในการเลือกคุณลักษณะนั้นสามารถสร้างแบบจำลองที่มีประสิทธิภาพในการจัดหมวดหมู่สถานที่ท่องเที่ยวสูงกว่าการเลือกคุณลักษณะโดยใช้เพียงแค่ชื่อของสถานที่ท่องเที่ยวเท่านั้นทั้งนี้แบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยวโดยเลือกคุณลักษณะจากทั้งชื่อและคำอธิบายของสถานที่ท่องเที่ยวจะมีประโยชน์อย่างยิ่งในการนำไปต่อยอดพัฒนาโปรแกรมประยุกต์หรือนำไปใช้เป็นแบบจำลองการจัดหมวดหมู่ข้อมูลสถานที่ท่องเที่ยวแหล่งอื่น ๆ อย่างเช่น วิกิพีเดีย ดีบีพีเดีย และจีโอเนมส์ หรือจากแหล่งสังคมออนไลน์อย่างเฟซบุ๊ค และโฟว์สแควร์ที่ในปัจจุบันได้รับความนิยมอย่างมาก

เอกสารอ้างอิง

Ali, W., Shamsuddin, S.M., and Ismail A.S. (2011). Web Proxy Cache Content Classification based on Support Vector Machine. **Journal of Artificial Intelligence** 4(1): 100-109.

Egger, R. and Buhalis, D. (2008). **eTourism Case Studies Management and Marketing Issues.** Oxford, UK: Butterworth-Heinemann.

Huang, Y., and Bian, L., (2009). A Bayesian network and analytic hierarchy process based personalized recommendations for tourist attractions over the Internet. **Expert Systems with Applications** 36: 933–943.

Jing, L.P., Huang, H.K., and Shi, H.B. (2002). Improved feature selection approach TFIDF in text mining. **ICMLA'02.Proceedings of the First International Conference on Machine Learning and Cybernetics** (pp. 944–946). Washington, DC, USA: IEEE Computer Society.

Kohavi.(1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. **IJCAI'95.Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2** (pp. 1137–1143). San Francisco: Morgan Kaufmann Publishers Inc.

Lee, C., and Lee, G.G. (2006).Information Gain and Divergence -base Feature Selection for Machine Learning-based Text Categorization. **Journal of Information Processing and Management: an International Journal - Special issue: Formal methods for information retrieval.** 42(1): 155-165.

Miller, T.W. & Nguyen H.T. (2005). **Data and Text Mining: A Business Applications Approach.** New York: Pearson Prentice Hall UpperSaddle River.

Mummidi, L., and Krumm, J. (2008). Discovering Points of Interest from Users' Map Annotations. **Journal of GeoJournal** 72 (3-4): 215-227.

Ozdikis, O., Orhan, F., and Danismaz, F. (2011). Ontology-based recommendation for points of interest retrieved from multiple data sources. **SWIM'11. Proceedings of the International Workshop on Semantic Web Information Management.**

Popescu, A., Grefenstette, G., and Moëllic, P-A.(2009a). Mining Tourist Information from User-Supplied Collections. **CIKM'09.Proceedings of the 18th ACM conference on Information and knowledge management** (pp. 1713-1716). NY, USA: ACM New York.

Popescu, A., Grefenstette, G. & Bouamor, H. (2009b).Mining a Multilingual Geographical Gazetteer from the Web. **WI-IAT'09. Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01** (pp. 58-65).Washington, DC, USA: IEEE Computer Society.

Sebastiani, F. (2005). **Text categorization**. In Alessandro Zanasi (ed.), Text Mining and its Applications, WIT Press, Southampton, UK, 2005, pp. 109-129.

Flickr. (2011). **Explore**[On-line]. Available: http://www.flickr.com/explore/

Google. (2011). **Google Maps** [On-line]. Available: http://maps.google.com/

Lonely Planet. (2010). **Resources** [On-line]. Available: http://developer.lplabs.com/index.php

Lonely Planet. (2011). **Thailand** [On-line]. Available: http://www.lonelyplanet.com/thailand

Wikipedia. (2011). **Siam Paragon** [On-line]. Available: http://en.wikipedia.org/wiki/Siam_Paragon