การจำแนกคุณภาพบทความวิกิพีเดียภาษาไทยโดยใช้ออนโทโลยี และลักษณะเด่นที่เกี่ยวกับการอ้างอิง

Quality Classification of Thai Wikipedia Articles using Ontology and Reference Feature Sets

กาญจนา แสงทองพัฒนา[1] และ นวลวรรณ สุนทรภิษัช[1*]

[1]ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ เขตจตุจักร กรุงเทพฯ 10903

[*]Corresponding Author, E-mail: fscinws@ku.ac.th

บทคัดย่อ

บทความวิกิพีเดียได้รับความนิยมและถูกใช้เป็นแหล่งอ้างอิงในเอกสารหรือสื่อต่าง ๆ วิกิพีเดียเป็นสารานุกรมเสรีที่อนุญาตให้ทุกคนสร้างและแก้ไขไบทความได้ ทำให้เกิดข้อกังวลเรื่องคุณภาพของบทความ นอกจากนี้ยังพบว่าบทความคุณภาพดีของวิกิพีเดียภาษาไทยมีจำนวนน้อย ดังนั้น งานวิจัยจึงนี้เสนอวิธีการจำแนกบทความวิกิพีเดียภาษาไทยตามคุณภาพสูงและต่ำ โดยมีสมมุติฐานคือ บทความคุณภาพสูงควรมีเนื้อหาที่ครอบคลุมคอนเซ็ปต์ส่วนใหญ่ในโดเมนนั้น ๆ รวมถึงควรมีความน่าเชื่อถือด้วย ผู้วิจัยได้ทดลองใช้ออนโทโลยีร่วมกับวิธีการเรียนรู้แบบต่าง ๆ พบว่าการทดลองใช้ออนโทโลยี ร่วมกับวิธีต้นไม้ตัดสินใจและลักษณะเด่นที่เกี่ยวกับการอ้างอิง สามารถจำแนกบทความที่มีคุณภาพสูงได้ค่า *F-Measure* เท่ากับ 0.85 สำหรับโดเมนบุคคล, 0.89 สำหรับโดเมนสัตว์, และ 0.72 สำหรับโดเมนสถานที่

ABSTRACT

Wikipedia articles are widely used as references sources in documents and other medias. Since the framework of Wikipedia allows user to create and edit articles, therefore the quality is the main concern. We found that there is small number of good quality in Thai articles, therefore this research proposes a method to classify Thai Wikipedia articles into high and low quality. The hypothesis is that the high quality articles should contain content that covers most concepts in the domain and should be reliable. We investigated ontology with various machine learning methods and found that using ontology and decision tree with reference features provides promising result measured in term of *F-Measure* which is 0.73 in biography domain, 0.89 in animal domain, and 0.72 in place domain.

คำสำคัญ: ออนโทโลยี วิกิพีเดียภาษาไทย ลักษณะเด่นแบบคอนเซ็ปต์ ลักษณะเด่นเกี่ยวกับการอ้างอิง ตัวจำแนกแบบเบย์อย่างง่าย ตัวจำแนกต้นไม้ตัดสินใจ

Keywords: Ontology, Thai Wikipedia, Concept Feature, Reference Feature, Naïve Bayes Classifier, Decision Tree Classifier

บทนำ

วิกิพีเดียเป็นสารานุกรมเสรีที่เปิดให้ผู้ใช้งานทั่วโลกสามารถใช้งานบทความได้โดยไม่มีค่าใช้จ่าย ผู้ใช้งานสามารถอ่าน สร้าง และปรับปรุงแก้ไขบทความได้อย่างเสรี ซึ่งมีข้อดีคือผู้ใช้งานสามารถช่วยกันพัฒนาเนื้อหาบทความได้ แต่อย่างไรก็ตาม พบว่าบางเนื้อหาอาจจะมีคุณภาพต่ำเนื่องจากยังไม่ได้รับการปรับปรุงแก้ไขให้ดีเพียงพอ วิกิพีเดียยังมีกลไกที่ช่วยกระตุ้นให้ผู้ใช้งานช่วยกันปรับปรุงแก้ไขบทความ กลไกหนึ่งคือผู้ใช้งานสามารถเสนอชื่อบทความคุณภาพสูงได้ ซึ่งบทความที่มีคุณภาพสูงของวิกิพีเดียภาษาไทยมี 2 ระดับคือ บทความคัดสรร (วิกิพีเดีย, 2558) และบทความคุณภาพ (วิกิพีเดีย, 2556) เมื่อผู้ใช้เสนอชื่อบทความคุณภาพสูง บทความนั้นจะถูกนำเข้าสู่กระบวนการพิจารณาและปรับปรุงบทความโดยผู้ใช้งานอื่นรวมถึงผู้ดูแลระบบภายในระยะเวลาสองสัปดาห์ หากเนื้อหาและองค์ประกอบของบทความตรงตามเกณฑ์ที่วิกิพีเดียกำหนด จะถูกประกาศชื่อให้เป็นบทความคัดสรร หรือบทความคุณภาพ จะเห็นได้ว่ากลไกดังกล่าวต้องอาศัยผู้ใช้งาน อนึ่ง บทความวิกิพีเดียภาษาไทยมีจำนวน 117,664 บทความ แม้จะมีผู้ใช้งานที่ลงทะเบียนทั้งหมด 292,503 คน แต่มีผู้ใช้งานที่มีการเคลื่อนไหวเพียง 1,038 คน ซึ่งน้อยกว่าจำนวนบทความถึงร้อยเท่า (วิกิพีเดีย, 2560) (ข้อมูล ณ วันที่ 6 มิ.ย. 2560) การสร้างวิธีการอัตโนมัติที่สามารถเสนอชื่อบทความคุณภาพสูงให้กับระบบวิกิพีเดียได้ จะช่วยให้กับผู้ใช้งานปรับปรุงแก้ไขบทความให้มีคุณภาพสูงได้สะดวกขึ้น และทำให้ระบบวิกิพีเดียมีบทความคุณภาพสูงจำนวนมากขึ้นด้วย

ผู้วิจัยเสนอวิธีการจำแนกคุณภาพบทความวิกิพีเดียภาษาไทยโดยอาศัยวิธีการทางออนโทโลยีร่วมกับวิธีการเรียนรู้แบบต่าง ๆ และพิจารณาคุณภาพบทความจากเนื้อหาซึ่งเป็นหนึ่งในคุณสมบัติของบทความคัดสรรและบทความคุณภาพตามเกณฑ์ของวิกิพีเดีย งานวิจัยนี้มีสมมุติฐานว่าการเขียนบทความวิกิพีเดียที่มีคุณภาพนั้นผู้เขียนควรเขียนบทความที่มีเนื้อหาครอบคลุมทุกประเด็น เนื้อหาภายใต้หัวข้อหลักควรมีคอนเซ็ปต์ที่ชัดเจน และบทความควรมีการอ้างแหล่งข้อมูลด้วย ซึ่งสอดคล้องกับเกณฑ์คุณภาพของบทความคัดสรรในวิกิพีเดีย แสดงในตารางที่ 1 อย่างไรก็ตามงานวิจัยนี้ไม่ได้พิจารณารูปภาพ เนื่องจากตามเกณฑ์ในวิกิพีเดีย บทความคุณภาพอาจมีรูปหรือไม่มีก็ได้

งานวิจัยที่เกี่ยวข้อง

หัวข้อนี้กล่าวถึงเกณฑ์มาตรฐานสำหรับบทความคุณภาพของวิกิพีเดีย งานวิจัยที่เกี่ยวข้องกับการจำแนกคุณภาพบทความวิกิพีเดีย และทฤษฎีที่เกี่ยวข้อง

1. มาตรฐานของการจำแนกบทความวิกิพีเดีย

วิกิพีเดียมีกระบวนการประเมินคุณภาพบทความโดยอาศัยผู้ใช้งานนำเสนอบทความให้กับระบบ วิกิพีเดียภาษาอังกฤษมีบทความทั้งหมด 5,430,603 บทความ แบ่งระดับคุณภาพเป็น 8 ระดับ (A, GA, B, C, Start, Stub, FL, and List) (Wikipedia, 2017) จำนวนบทความวิกิพีเดียภาษาอังกฤษมีมากกว่าภาษาไทยถึง 46 เท่า จึงเป็นเหตุผลหนึ่งที่วิกิพีเดียภาษาไทยได้แบ่งบทความคุณภาพสูงเป็นสองประเภทเท่านั้นคือ บทความคัดสรรและบทความคุณภาพ

กระบวนการประเมินคุณภาพบทความเริ่มต้นด้วยการเสนอชื่อบทความ หลังจากนั้นภายในสองสัปดาห์เป็นกระบวนการการสนับสนุนและคัดค้าน ซึ่งเป็นขั้นตอนสำคัญที่ผู้ใช้งานจะช่วยกันปรับปรุงบทความจนกว่าจะได้รับการสนับสนุน หากสิ้นสุดขั้นตอนดังกล่าวแต่บทความยังอยู่ระหว่างการปรับปรุงหรือบทความถูกคัดค้าน ผู้ใช้งานสามารถเสนอชื่อบทความเข้ามาใหม่ได้อีก เมื่อแก้ไขบทความให้มีคุณภาพเหมาะสมแล้ว หากบทความได้รับการสนับสนุน ผู้ดูแลระบบจะทำการประกาศชื่อเป็นบทความคุณภาพสูงต่อไป ซึ่งเป็นบทความประเภทหนึ่งที่ผู้อ่านสามารถนำไปใช้ประโยชน์ได้ด้วยความเชื่อมั่นในเนื้อหาและแหล่งอ้างอิงของบทความนั้น อย่างไรก็ตาม บทความวิกิพีเดียภาษาไทยมีบทความคัดสรรจำนวน 138 บทความ และบทความคุณภาพจำนวน 136 บทความ คิดเป็นเพียงร้อยละ 0.2 ของบทความทั้งหมดเท่านั้น (ข้อมูล ณ วันที่ 6 กรกฎาคม 2560 จาก

https://th.wikipedia.org/wiki/วิกิพีเดีย:บทความคัดสรร และ https://th.wikipedia.org/wiki/วิกิพีเดีย:บทความคุณภาพ)

บทความคุณภาพมีเกณฑ์คุณภาพเป็นรองบทความคัดสรร บทความคัดสรรเป็นบทความที่ดีที่สุด บทความคุณภาพเป็นบทความที่เขียนดีน่าพอใจ (วิกิพีเดีย, 2559) หากบทความคุณภาพถูกพัฒนาเนื้อหา ความเป็นกลาง และรูปแบบให้เป็นไปตามเกณฑ์ของบทความคัดสรร บทความคุณภาพดังกล่าวจะถูกเพิ่มระดับเป็นบทความคัดสรรได้ อย่างไรก็ตาม ผู้ใช้งานสามารถเสนอให้มีการปรับปรุงข้อบกพร่องหรือลดระดับของบทความคัดสรร หรือบทความคุณภาพออกจากรายการบทความดังกล่าวได้ด้วย เกณฑ์มาตรฐานสำหรับบทความคุณภาพสูงของวิกิพีเดียแสดงได้ดังตารางที่ 1 (Wikimedia Foundation., 2017) พบว่าบทความคุณภาพเขียนเนื้อเรื่องกว้าง ๆ แต่บทความคัดสรรมีเนื้อหาละเอียดครอบคลุม ในงานวิจัยนี้กำหนดคุณภาพบทความเป็น 2 ระดับคือบทความคุณภาพสูง และบทความคุณภาพต่ำ โดยมุ่งประเด็นที่เนื้อหาบทความซึ่งครอบคลุมทุกคอนเซ็ปต์ของโดเมนนั้น ๆ และนำเสนอในรูปแบบฐานความรู้ออนโทโลยี นอกจากนี้ยังได้พิจารณาองค์ประกอบของบทความตามเกณฑ์วิกิพีเดียด้วย คือ หัวข้อ และแหล่งอ้างอิง

ตารางที่ 1 คุณสมบัติของบทความคัดสรร และบทความคุณภาพ

คุณสมบัติ	บทความคัดสรร	บทความคุณภาพ
อธิบายเนื้อหาดี มีแหล่งอ้างอิงที่เชื่อถือได้ เป็นกลาง มีเสถียรภาพ	/	/
เนื้อหาครอบคลุมทุกประเด็น	/	
มีการจัดเรียงหัวข้อตามรูปแบบพื้นฐานของวิกิพีเดีย	/	/
มีการจัดเรียงหัวข้อตามตามความสำคัญ หรือตามลำดับเวลา	/	
เนื้อหากว้าง ไม่กล่าวถึงรายละเอียดที่ไม่จำเป็น	/	/
อาจมีบทความแยกย่อยถ้าจำเป็น (หัวข้อ)	/	
ต้องมีรูปภาพที่ไม่ละเมิดลิขสิทธิ์และมีการระบุแหล่งที่มา	/	
ไม่จำเป็นต้องมีรูปภาพ หากมี ต้องไม่ละเมิดลิขสิทธิ์และมีการระบุแหล่งที่มา		/

2. งานวิจัยที่เกี่ยวข้องกับการจำแนกคุณภาพบทความวิกิพีเดีย

จากการศึกษาพบว่าในวิกิพีเดียภาษาอังกฤษมีการพัฒนาเครื่องมือ WikiLyzer เพื่อช่วยงานต่าง ๆ ในการประเมินคุณภาพบทความ เครื่องมือนี้ช่วยผู้ใช้ในการค้นหาบทความที่ดีและระบุจุดอ่อนเพื่อการปรับปรุงบทความต่อไปได้ (Sciascio et al., 2017) ก่อนหน้านี้ GreenWiki (Dalip et al., 2011) ถูกพัฒนาขึ้นเพื่อประเมินคุณภาพบทความและแสดงตัวบ่งชี้ที่ช่วยให้ผู้ใช้ทราบถึงข้อสรุปของคุณภาพบทความของเขา ในปี 2010 WikipediaViz เป็นเครื่องมือที่ถูกพัฒนาขึ้นมาให้มีรูปแบบการนำเสนอเป็นภาพ (Visualization tools) (Chevalier et al., 2010) ซึ่งถูกสร้างด้วยภาษา PHP และรวมเข้ากับระบบมีเดียวิกิ (Mediawiki system) งานวิจัยนี้เสนอตัวชี้วัด 5 แบบสำหรับการประเมินความครบถ้วนและคุณภาพบทความวิกิพีเดียโดยนำเสนอเป็นภาพเพื่อให้ผู้อ่านมีส่วนร่วมในการแก้ไขบทความด้วย ซึ่งพบว่าช่วยลดเวลาในการประเมินคุณภาพบทความได้ และยังรักษาความมีคุณภาพของบทความไว้ได้ด้วย

งานวิจัยหลายชิ้นได้ศึกษาชุดลักษณะเด่นของบทความโดยอาศัยคำสำคัญ การอ้างอิง และการเชื่อมโยงระหว่างหน้าเว็บเป็นต้น เช่นการใช้ลิงค์ภายนอก (External links) (Tzekou et al., 2011) พบว่าหากระบบวิกิพีเดียมีการกลไกที่ดีในการจัดการเรื่องการเชื่อมโยงไปยังข้อมูลภายนอก จะลดจำนวนการเชื่อมโยงไปยังข้อมูลที่ไม่มีจริงได้ และภาพรวมของคุณภาพบทความจะดีขึ้น

วิธีการเรียนรู้ด้วยเครื่องจักร ถูกนำมาใช้ในหลายงานวิจัยเพื่อหาลักษณะเด่นที่สามารถใช้ในการจำแนกคุณภาพบทความวิกิพีเดียได้ ตัวอย่างเช่น ลักษณะเด่นแบบข้อความ ลักษณะเด่นแบบเมทาดาทา (Meta data features) หรือการอธิบายข้อมูลในรูปแบบโครงสร้าง (Javanmardi, 2011) นอกจากนี้ ยังมีการนำความยาว การอ้างอิง การเชื่อมโยงไปยังบทความอื่น และรูปภาพ (Calzada and Dekhtyar, 2010) มาเป็นลักษณะเด่นในการจำแนกคุณภาพบทความด้วย และต่อมามีงานวิจัยที่ใช้ลักษณะเด่นดังกล่าวร่วมกับคำสำคัญของโดเมนที่ปรากฏในบทความ (Saengthongpattana and Soonthorn-

phisaj, 2014) ซึ่งพบว่ามีนัยสำคัญในการจำแนกคุณภาพบทความได้

การคัดเลือกลักษณะเด่นเป็นงานที่มีความยาก เนื่องจากแต่ละลักษณะเด่นมีข้อดีและข้อด้อยที่แตกต่างกัน ด้วยเหตุนี้ มีงานวิจัยที่ไม่ใช้ลักษณะเด่นของบทความ (Dang and Ignat, 2016) แต่ประยุกต์ใช้กระบวนการในการตรวจสอบบทความของผู้ประเมินบทความ (Reviewers) โดยสร้างชุดข้อมูลนำเข้าในรูปแบบเวกเตอร์ และอาศัยวิธีการแบบโครงข่ายประสาทเทียมเชิงลึก (Deep neural network) ซึ่งวิธีการดังกล่าวให้ค่าความถูกต้องสูง นอกจากนี้ยังพบงานวิจัยสำหรับวิกิพีเดียภาษาสเปน (Betancourt et al., 2016) ซึ่งมีการศึกษาลักษณะเฉพาะของการทำงานในทีมบรรณาธิการ และทดลองบนอัลกอริทึมต้นไม้ตัดสินใจ งานวิจัยดังกล่าวแสดงให้เห็นว่าลักษณะเฉพาะที่ได้จากทีมบรรณาธิการให้ประสิทธิภาพสูงในการทำนาย และระยะเวลาการมีส่วนร่วมทั้งหมดเป็นตัวชี้วัดที่มีความสำคัญมากที่สุด

จะเห็นได้ว่าบทบาทของบรรณาธิการมีความสำคัญต่อการปรับปรุงคุณภาพบทความ Li และคณะ (Li et al., 2015) จึงพัฒนาหลายโมเดลเพื่อจัดลำดับบทความโดยใช้ความสัมพันธ์ระหว่างการแก้ไขปรับปรุงบทความกับบรรณาธิการ และพบว่าการใช้คู่มือประเมินคุณภาพ (Manual evaluation) เพื่อช่วยประเมินแบบอัตโนมัติ เป็นทางออกที่ดีสำหรับงานประเมินคุณภาพบทความวิกิพีเดีย นอกจากนี้ ในปีที่ผ่านมาพบงานวิจัย (Suzuki and Nakamura, 2016) ที่ศึกษาบทบาทการมีส่วนร่วมของบรรณาธิการเป็นปัจจัยหนึ่งที่เกี่ยวกับคุณภาพของบทความ เพราะบรรณาธิการที่ดีจะสามารถปรับปรุงคุณภาพบทความได้ ดังนั้น งานวิจัยดังกล่าวจึงเสนอวิธีการประเมินคุณภาพของบรรณาธิการวิกิพีเดีย โดยใช้กระบวนการร่วมสร้างสรรค์ของกลุ่มคน (Crowdsourcing) เพื่อการตรวจจับการเปลี่ยนความหมายของข้อความด้วยตัวเอง

3. ตัวจำแนกแบบเบย์อย่างง่าย

ตัวจำแนกแบบเบย์อย่างง่าย ใช้หลักความน่าจะเป็น มีกระบวนการทำงานที่ไม่ซับซ้อนและใช้เวลาไม่นาน จึงเหมาะสมในการประยุกต์ใช้งานหลายด้าน เช่น การแยกแยะประเภทวัตถุ และการจำแนกประเภทเอกสาร โดยใช้ตัวจำแนกแบบเบย์อย่างง่ายคำนวณความน่าจะเป็นแบบมีเงื่อนไขของประเภทบทความที่ต้องการจำแนก โดยให้ T เป็นชุดข้อมูลที่ใช้ฝึก ประกอบด้วย M ตัวอย่างสำหรับการฝึกคือ $\{(X_1, y_1),...,(X_m, y_m)\}$ ขณะที่ X_i คือเวกเตอร์ของลักษณะเด่นแทนด้วย $\{x_{i,1},...,x_{i,n}\}$ (n คือจำนวนลักษณะเด่น) และ $y_k \in \{c_1,...,c_l\}$ ซึ่ง c_l คือเซ็ตของประเภทบทความ

อัลกอริทึมสร้างโมเดลสำหรับตัวจำแนกเป็นสมมุติฐาน $y = f(X)$ ตัวจำแนกจะทำการทำนายประเภทโดย $y_t \in \{c_1,...,c_l\}$ สำหรับ (X_t, y_t) คือข้อมูล X_t, ที่ต้องการทำนายประเภท y_t โดยคำนวณจาก

$$y_t = \text{argmax}_{y_k} \left(\text{P}(y_k)\right) \prod_{j=1}^{N} \text{P}(x_j|y_k) \quad (1)$$

ความน่าจะเป็นก่อนหน้า $\text{P}(y_k)$ คือการประมาณอัตราระหว่างจำนวนของข้อมูลตัวอย่างในประเภท y_k กับจำนวนตัวอย่างทั้งหมด ความน่าจะเป็นแบบมีเงื่อนไข คือความน่าจะเป็นของการพบ x_j ปรากฏในประเภท y_k

ตัวอย่างงานวิจัยที่นำตัวจำแนกแบบเบย์อย่างง่ายไปใช้งานคือ (Harpalani et al., 2010) การหาว่าเกิดการก่อกวน (Vandalism) ขึ้นในการแก้ไขบทความหรือไม่ โดยอาศัยข้อมูลที่เกี่ยวข้องกับบรรณาธิการ ข้อคิดเห็นในการแก้ไขข้อมูล คุณภาพของการแก้ไข เช่นการสะกดผิด และรูปแบบทั่วไปในการก่อกวน

4. ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ เป็นโมเดลทางคณิตศาสตร์ที่ใช้ทำนายประเภทของวัตถุโดยพิจารณาจากลักษณะของวัตถุ โหนดภายในของต้นไม้จะแสดงตัวแปร ส่วนกิ่งจะแสดงค่าที่เป็นไปได้ของตัวแปร ส่วนโหนดใบจะแสดงประเภทของวัตถุ การสร้างต้นไม้ตัดสินใจอาศัยค่าเกนมาตรฐาน (Information gain) เป็นตัวคัดเลือกโหนด โดยหาค่าเกนมาตรฐานทุก ๆ ลักษณะเด่นในข้อมูล แล้วเลือกโหนดที่มีค่าเกนมาตรฐานมากที่สุดมาเป็นโหนดเริ่มต้น โดยค่าเกนมาตรฐาน (Gain(A)) คำนวณจากการหาค่าเอนโทรปีก่อนการแบ่งแยก (Entropy(D)) ลบด้วยค่าเอนโทรปีหลังแบ่งแยกด้วยลักษณะเด่นที่พิจารณา ($\text{Entropy}_A(D)$)

โดยมีค่าสมการคือ

$$\text{Entropy(D)}_t = -\sum_{i=1}^{n} P_i \, log \, P_i \quad (2)$$

$$\text{Entropy}_A\text{(D)} = -\sum_{i=1}^{n} \frac{|D_j|}{|D|} \times \text{Entropy (D}_j\text{)} \quad (3)$$

5. แบบจำลองเวกเตอร์สเปซ และการให้น้ำหนักคำ

แบบจำลองเวกเตอร์สเปซ และการให้น้ำหนักคำ (Vijay and S. K. M., 1986) (Jain and Li, 1998) เป็นหนึ่งในวิธีการแทนเอกสารที่ไม่มีโครงสร้างด้วยแบบจำลองเวกเตอร์สเปซ โดยกำหนดให้เอกสารแต่ละฉบับเปรียบเสมือนเวกเตอร์ของคำ ขนาดของเวกเตอร์ขึ้นอยู่กับจำนวนของคำที่ปรากฏอยู่ในเอกสารฉบับนั้นกำหนดให้ W_{ik} คือน้ำหนักของคำ K ที่ปรากฏในเอกสารฉบับที่ i เวกเตอร์สำหรับเอกสาร D_i เขียนแทนด้วย D_i = (W_{i1}, W_{i2},...,W_{it}) ซึ่ง t คือจำนวนของคำที่ไม่ซ้ำกัน ในชุดของเอกสารทั้งหมด ดังนั้นในสเปซของเอกสารชุดหนึ่งจะมีมิติเท่ากับ t-มิติ

6. ออนโทโลยี

ออนโทโลยีเป็นโครงสร้างข้อมูลที่นำมาประยุกต์ใช้แทนคอนเซปต์ของบทความ ประกอบด้วยโหนดและความสัมพันธ์ ซึ่งโครงสร้างความสัมพันธ์ดังกล่าวเครื่องคอมพิวเตอร์สามารถเข้าใจและแปลความได้จากภาษาที่ใช้ในออนโทโลยีเพื่อบรรยายข้อมูลเชิงความหมาย ได้แก่ XML (Extensible Markup Language) RDF (Resource Description Framework) และ OWL (Web Ontology Language)

เครื่องมือที่สนับสนุนการพัฒนาออนโทโลยีในปัจจุบันที่ได้รับความนิยม เช่น โปรแกรม Protege พัฒนาโดยมหาวิทยาลัยสแตนฟอร์ด (http://protege.stanford.edu/) และโปรแกรม Hozo ซึ่งพัฒนาโดยมหาวิทยาลัยโอซากา (http://www.hozo.jp/) เป็นต้น เครื่องมือเหล่านี้ช่วยให้วิศวกรความรู้ หรือผู้เชี่ยวชาญเฉพาะสาขา สามารถถ่ายทอดและจัดเก็บองค์ความรู้ในรูปแบบของออนโทโลยีได้สะดวกยิ่งขึ้น

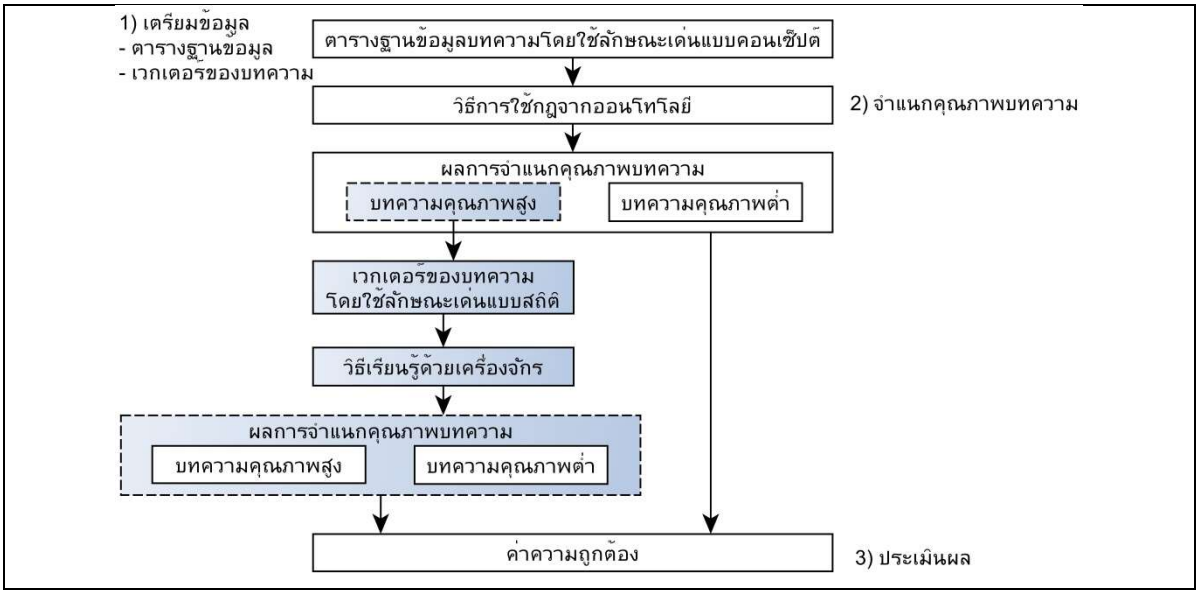
ฐานความรู้ออนโทโลยีที่เป็นที่รู้จักโดยทั่วไปคือ DBpedia (http://wiki.dbpedia.org) มีที่มาจากชุมชนข้อมูลขนาดใหญ่ (crowd-sourced community) พยายามสร้างข้อมูลที่มีโครงสร้างจากวิกิพีเดีย และให้ข้อมูลใช้ได้บนเว็บด้วย DBpedia มีการอธิบายข้อมูลถึง 4 ล้านสิ่ง เช่น คน สถานที่ องค์กร สายพันธุ์ และโรคต่าง ๆ นอกจากนี้ มีหลายงานวิจัย (Hartig, 2013) (Pan et al., 2006) พยายามสร้างกลไกในการค้นคืนข้อมูลจากฐานความรู้เชิงความหมาย (Semantic knowledge bases)

นอกจากงานวิจัยที่อาศัยข้อมูลวิกิพีเดียภาษาอังกฤษแล้ว ฐานความรู้ออนโทโลยียังถูกใช้งานเพื่อนำเสนอความรู้ในภาษาอาราบิกด้วย (Nora et al., 2011) งานวิจัยดังกล่าวสร้างออนโทโลยีเพื่อนำเสนอคอนเซ็ปต์และข้อมูลในเชิงความหมายที่เกี่ยวข้องสำหรับโดเมนนั้น ๆ โมเดลออนโทโลยีสร้างจากข้อมูลวิกิพีเดียอาราบิก โดยสกัดความสัมพันธ์เชิงความหมายจากกล่องข้อมูลและรายการของประเภทบทความ โมเดลที่ได้แสดงให้เห็นว่าข้อมูลที่ได้รับนั้นมีความสอดคล้องกัน

สำหรับภาษาที่อยู่ในภูมิภาคอาเซีย ได้มีงานวิจัยที่อาศัยข้อมูลวิกิพีเดียภาษาญี่ปุ่นเพื่อสร้างออนโทโลยีภาษาญี่ปุ่นขนาดใหญ่ (Tamagawa et al., 2010) วิกิพีเดียถูกใช้เป็นทรัพยากรที่มีคุณค่าสำหรับการเรียนรู้แบบออนโทโลยี โดยมุ่งเน้นที่ความสัมพันธ์แบบจัดเป็น (“is-a”) นอกจากนี้ก็ยังความสัมพันธ์แบบอื่นด้วย พบว่าคำที่มีความหมายเดียวกัน (Hyponym) สามารถสกัดได้โดยใช้ความสัมพันธ์ในออนโทโลยี

วิธีการดำเนินการวิจัย

วิธีการดำเนินการวิจัยประกอบด้วย 3 ขั้นตอนหลักคือ 1) การเตรียมข้อมูลเพื่อใช้ในการทดลอง 2) การจำแนกคุณภาพบทความด้วยวิธีออนโทโลยี และวิธีการเรียนรู้ด้วยเครื่องจักร เพื่อแสดงผลว่าบทความมีคุณภาพสูงหรือต่ำ และ 3) การประเมินผลเพื่อให้ทราบว่าคุณภาพของบทความที่ถูกจำแนกมีความถูกต้องอย่างไร แสดงได้ดังรูปที่ 1



รูปที่ 1 วิธีการดำเนินการวิจัย

1. การเตรียมข้อมูล

เบื้องหลังของบทความวิกิพีเดียที่มีรูปแบบสวยงามบนหน้าเว็บเพจนั้น คือบทความที่มีโครงสร้างประกอบด้วยวิกิแทกและสัญญลักษณ์กำกับข้อความเพื่อบ่งบอกหน้าที่ของข้อความนั้น เช่น <title>เนลสัน มันเดลา</title> หมายถึงชื่อบทความ เนลสัน มันเดลา == ช่วงแรกของชีวิต == หมายถึง หัวข้อหลัก ชื่อช่วงแรกของชีวิต เป็นต้น ตัวอย่างโครงสร้างของบทความวิกิพีเดียแสดงได้ดังรูปที่ 2

หลังจากที่ได้ทำการบันทึกชุดข้อมูลบทความจากเว็บไซต์วิกิพีเดียภาษาไทยแล้ว ก่อนเข้าสู่ขั้นตอนการเตรียมข้อมูล ทุกบทความจะต้องผ่านการตัดคำแบบเลือกคำที่ยาวที่สุด (Longest Matching) การเตรียมข้อมูลประกอบด้วย 5 ขั้นตอน คือ

(1) การแบ่งชุดข้อมูลตามโดเมน (2) การระบุคุณภาพให้กับชุดข้อมูล (3) การกำหนดคอนเซ็ปต์ และคำสำคัญ (4) การสร้างตารางฐานข้อมูล เพื่อเก็บบทความโดยใช้ลักษณะเด่นแบบคอนเซ็ปต์ และ (5) การสร้างเวกเตอร์ของบทความโดยใช้ลักษณะเด่นแบบสถิติ แสดงได้ดังรูปที่ 3

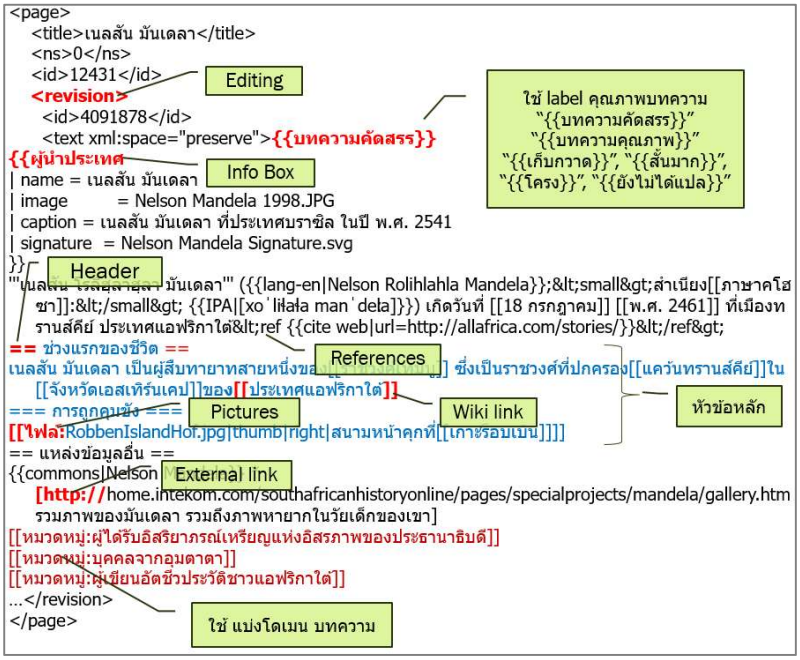
1.1 การแบ่งชุดข้อมูลตามโดเมน

ชุดข้อมูลที่ใช้ในการวิจัยนี้บันทึกมาจากบทความวิกิพีเดียภาษาไทยจำนวน 14,996 บทความ ประเภทของโดเมนถูกกำหนดโดยผู้เขียนบทความ ซึ่งปรากฏที่ส่วนท้ายของแต่ละบทความโดยอยู่หลังข้อความ “[[หมวดหมู่:” (แสดงในรูปที่ 2) งานวิจัยนี้ใช้โดเมนที่มีจำนวนบทความสูงที่สุดสามลำดับแรกคือ โดเมนบุคคล โดเมนสถานที่ และโดเมนสัตว์ ซึ่งโดเมนเหล่านี้มีทั้งบทความคุณภาพสูงและคุณภาพต่ำ

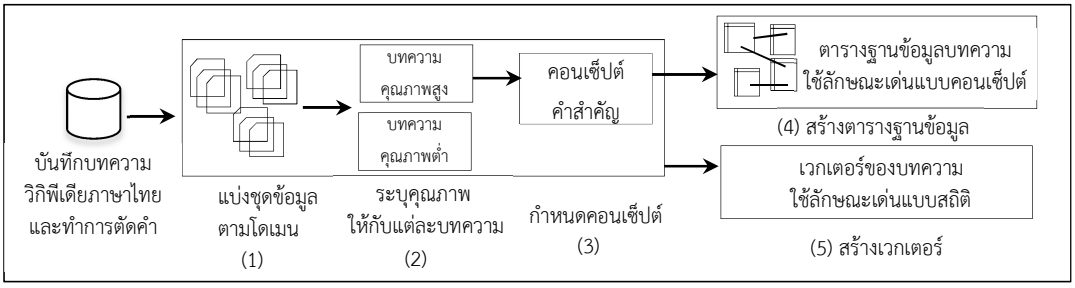
1.2 การระบุคุณภาพให้กับชุดข้อมูล

สำหรับการระบุคุณภาพให้กับแต่ละบทความ พบว่าในหน้าเว็บเพจของบทความคัดสรรจะปรากฏสัญลักษณ์รูปดาว () และในหน้าของบทความคุณภาพจะปรากฏสัญลักษณ์รูปเครื่องหมายบวกอยู่ในวงกลม () ซึ่งสัญลักษณ์ดังกล่าวจะปรากฏที่ส่วนบนสุดด้านขวามือของหน้าเว็บเพจบทความ สำหรับหน้าของบทความที่ต้องแก้ไขปรับปรุง จะปรากฏสัญลักษณ์รูปไม้กวาด () ทั้งนี้ ตามโครงสร้างของบทความวิกิพีเดีย สามารถพิจารณาได้จาก {{บทความคัดสรร}} {{บทความคุณภาพ}} {{เก็บกวาด}} {{สั้นมาก}} {{ยังไม่ได้แปล}} เป็นต้น (แสดงในรูปที่ 2)

สำหรับงานวิจัยนี้แบ่งคุณภาพของบทความเป็น 2 ประเภทคือ (1) บทความคุณภาพสูง ซึ่งได้จากบทความคัดสรรและบทความคุณภาพ และ (2) บทความคุณภาพต่ำ ซึ่งได้จากบทความที่ต้องแก้ไขปรับปรุง จำนวนบทความที่ใช้ในงานวิจัยนี้แสดงได้ดังตารางที่ 2



รูปที่ 2 ตัวอย่างโครงสร้างของบทความวิกิพีเดีย



รูปที่ 3 การเตรียมข้อมูล

ตารางที่ 2 จำนวนบทความที่ใช้ในงานวิจัยนี้

ชื่อโดเมน	จำนวนบทความ	
	คุณภาพสูง	คุณภาพต่ำ
บุคคล	66	8,596
สถานที่	27	6,043
สัตว์	9	255
จำนวนรวม	102	14,894

1.3 การกำหนดคอนเซ็ปต์

สมมุติฐานประการหนึ่งของงานวิจัยนี้คือ บทความคุณภาพสูงควรมีเนื้อหาที่ครอบคลุมทุกคอนเซ็ปต์ของโดเมนนั้น และเนื้อหาในแต่ละหัวข้อหลักของบทความควรมีเพียงหนึ่งคอนเซ็ปต์เท่านั้น ในแต่ละคอนเซ็ปต์ประกอบด้วยคำสำคัญที่เป็นตัวแทนของคอนเซ็ปต์ได้ เช่น คอนเซ็ปต์การศึกษาประกอบด้วยคำสำคัญเช่น เข้าเรียนต่อ จบปริญญา และระดับประกาศนียบัตร เป็นต้น คอนเซ็ปต์การทำงาน ประกอบด้วยคำสำคัญเช่น ดำรงตำแหน่ง เลื่อนยศ ได้รับการแต่งตั้ง เป็นต้น

งานวิจัยนี้อาศัยผู้เชี่ยวชาญสร้างคอนเซ็ปต์และใช้วิธีกึ่งอัตโนมัติในการหาคำสำคัญของแต่ละคอนเซ็ปต์ โดยพิจารณาจากเนื้อหาที่สอดคล้องกันของบทความคุณภาพสูงที่อยู่ในโดเมนเดียวกัน พบว่าบทความในโดเมนบุคคล ควรประกอบด้วยคอนเซ็ปต์เกี่ยวกับการเกิด การศึกษา อาชีพ ความเกี่ยวข้องกับบุคคล

อื่น รวมถึงอาจมีเนื้อหาเกี่ยวกับการเสียชีวิต บทความในโดเมนสถานที่ควรมีคอนเซ็ปต์เกี่ยวกับพิกัดที่ตั้ง ยุคที่สร้าง จุดประสงค์ในการสร้าง รวมถึงการดูแลจัดการด้วย บทความในโดเมนสัตว์ควรประกอบด้วยคอนเซ็ปต์เกี่ยวกับรูปร่าง ลักษณะการเกิด สายพันธุ์ อาหาร และถิ่นที่อยู่ รายการคอนเซ็ปต์และตัวอย่างคำสำคัญแสดงในตารางที่ 3

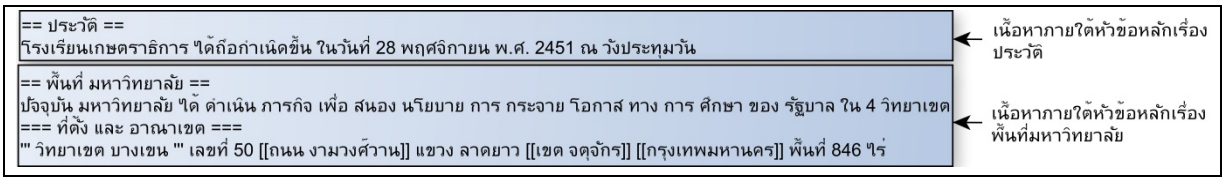
แต่ละโดเมนประกอบด้วยคอนเซ็ปต์ ในแต่ละคอนเซ็ปต์ประกอบด้วยคำสำคัญ คำสำคัญทุกคำในแต่ละโดเมนจะถูกใช้เป็นลักษณะเด่นในการสร้างเวกเตอร์ของชุดข้อมูลในแต่ละโดเมน ซึ่งจะกล่าวต่อไปในหัวข้อการสร้างลักษณะเด่นแบบสถิติ

ตารางที่ 3 รายการคอนเซ็ปต์และตัวอย่างคำสำคัญของคอนเซ็ปต์ในแต่ละโดเมน

ลำดับ	คอนเซ็ปต์	คำอธิบาย	ตัวอย่างคำสำคัญ
โดเมนบุคคล			
1	การเกิด	ข้อมูลวันเกิดและสถานที่เกิด	เกิดวันที่, เกิดที่, สถานที่เกิด
2	การศึกษา	ข้อมูลการศึกษา การฝึกอบรม ความรู้ความสามารถ	จบการศึกษา, ศึกษาสาขา, เชี่ยวชาญด้าน
3	การทำงาน	ข้อมูลการประกอบอาชีพ ตำแหน่งหน้าที่การงาน	ดำรงตำแหน่ง, เลื่อนยศ, ได้รับการแต่งตั้ง
4	ความเกี่ยวข้องกับบุคคลอื่น	ความสัมพันธ์ ความรู้จักกับบุคคลอื่น เช่น พ่อแม่ ญาติ	มีบุตร, อภิเษกสมรสกับ, ลุงของ, เพื่อน
5	การเสียชีวิต	ข้อมูลการเสียชีวิต เช่น วันที่ สาเหตุ อายุ	สิ้นชีพ, ถูกสังหาร, อายุรวม
โดเมนสถานที่			
1	ที่ตั้ง	ตำแหน่งที่ตั้ง พิกัด บริเวณ	พื้นที่ใกล้, ตั้งอยู่บน, พิกัด
2	ยุคที่สร้าง	ข้อมูลการก่อตั้ง ยุค ช่วงเวลา	ยุคแรก, ก่อตั้งเมื่อ, ช่วงทศวรรษ
3	จุดประสงค์	จุดประสงค์ในการสร้าง	สร้างขึ้นเพื่อ, การใช้งาน, เป้าหมาย
4	การดูแลรักษา	ข้อมูลการดูแลรักษาสถานที่ การบริหารจัดการ	ดำเนินการโดย, งบประมาณ, องค์การบริหาร
โดเมนสัตว์			
1	รูปร่างลักษณะ	ข้อมูลขนาด รูปร่าง ส่วนประกอบที่มองเห็นได้	กะโหลก, สีขน, ขนาดและรูปร่าง
2	สายพันธุ์	ข้อมูลต้นตระกูล และสายพันธุ์	ต้นตระกูล, สปีชีส์, เครือญาติ
3	การขยายพันธุ์	ข้อมูลการขยายพันธุ์ การสืบพันธุ์ การให้กำเนิด	ข้ามสายพันธุ์, วางไข่, ออกลูกเป็นตัว
4	อาหาร	ข้อมูลอาหาร พฤติกรรมการล่าเหยื่อ	จับปลา, ล่าเหยื่อ, สัตว์กินเนื้อ
5	ถิ่นที่อยู่	ข้อมูลพื้นที่ที่พบสัตว์ชนิดนั้น ที่อยู่อาศัยของสัตว์	ภายในป่า, ถิ่นที่อยู่, อาณาเขต

1.4 การสร้างตารางฐานข้อมูล

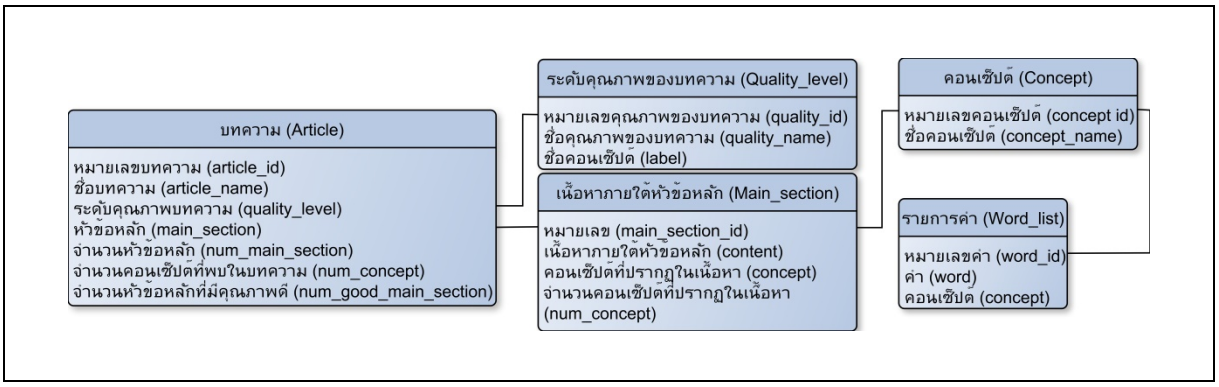
งานวิจัยนี้ให้ความสำคัญกับเนื้อหาภายใต้หัวข้อหลัก บทความวิกิพีเดียใช้เครื่องหมายเท่ากับสองตัว (==) นำหน้าและปิดท้ายชื่อหัวข้อหลัก และใช้เครื่องหมายเท่ากับสามตัว (===) นำหน้าและปิดท้ายชื่อหัวข้อย่อยภายใต้หัวข้อหลักนั้น ตัวอย่างบทความในรูปที่ 4 ประกอบด้วย 2 รายการ คือเนื้อหาภายใต้หัวข้อหลักเรื่องประวัติ และเนื้อหาภายใต้หัวข้อหลักเรื่องพื้นที่มหาวิทยาลัย ทั้งนี้ ข้อมูลที่นำมาใช้จะถูกตัดคำก่อน



รูปที่ 4 ตัวอย่างเนื้อหาภายใต้หัวข้อหลักของบทความวิกิพีเดีย

ฐานข้อมูลประกอบด้วย 5 ตารางหลักที่มีความสัมพันธ์กันคือ ตารางบทความ (Article) ตารางเนื้อหาภายใต้หัวข้อหลัก (Main_section) ตารางคอนเซ็ปต์ (Concept) ตารางระดับคุณภาพของบทความ (Quality_level) และตารางรายการคำ (Word_list) ในแต่ละตารางประกอบด้วยคอลัมน์ซึ่งเป็นองค์ประกอบของบทความ แสดงความสัมพันธ์ของตารางฐานข้อมูลและคอลัมน์ได้ดังรูปที่ 5

รูปที่ 5 แสดงความสัมพันธ์ระหว่างตาราง กล่าวคือ ตารางบทความประกอบด้วย คอลัมน์ระดับคุณภาพบทความเป็นการจัดเก็บหมายเลขระดับคุณภาพบทความเพื่อใช้เชื่อมโยงกับตารางระดับคุณภาพบทความ คอลัมน์หัวข้อหลักเป็นการจัดเก็บหมายเลขหัวข้อหลักเพื่อใช้เชื่อมโยงกับตารางเนื้อหาภายใต้หัวข้อหลัก เป็นต้น สำหรับตารางหัวข้อหลัก ประกอบด้วยคอลัมน์จำนวนคอนเซ็ปต์ที่พบในเนื้อหา เป็นการจัดเก็บหมายเลขคอนเซ็ปต์เพื่อใช้เชื่อมโยงกับตารางคอนเซ็ปต์ และสำหรับตารางคอนเซ็ปต์ ประกอบด้วยคอลัมน์ชื่อคอนเซ็ปต์ ในส่วนนี้สามารถเชื่อมโยงกับคำสำคัญของคอนเซ็ปต์ได้ ซึ่งคำสำคัญถูกสร้างขึ้นมาในขั้นตอนการกำหนดคอนเซ็ปต์แล้ว โดยเก็บไว้เป็นรายการคำสำหรับแต่ละครเซ็ปต์ของแต่ละโดเมน



รูปที่ 5 ความสัมพันธ์ของตารางฐานข้อมูลและคอลัมน์ซึ่งเป็นองค์ประกอบของแต่ละตาราง

1.5 การสร้างลักษณะเด่นแบบสถิติ

งานวิจัยนี้ได้ทดลองและพบว่าการเชื่อมโยงไปยังบทความอื่นของวิกิพีเดีย การเชื่อมโยงไปยังข้อมูลภายนอก การอ้างอิงแหล่งที่มา และความยาวของบทความ เป็นคุณสมบัติหนึ่งซึ่งอยู่ในเกณฑ์คุณภาพของบทความวิกิพีเดีย (Soonthornphisaj and Paengporn, 2017) (แสดงไว้ในตารางที่ 4) ผู้วิจัยได้สร้างเว็กเตอร์ที่ประกอบด้วยจำนวนของลักษณะเด่นที่ปรากฏในบทความ โดยขนาดของเวกเตอร์เท่ากับจำนวนของลักษณะเด่น ตัวอย่างเช่น เว็กเตอร์ของบทความเรื่อง “เจ. เค. โรว์ลิง” = (235, 226, 23.32743, 289, 218, 187746) หมายถึงบทความนี้มีจำนวนการเชื่อมโยงวิกิ 235, จำนวนการเชื่อมโยงเว็บเพจ 226, จำนวนเฉลี่ยของการเชื่อมโยงโดเมนเนม 23.32743, จำนวนครั้งของการอ้างอิง 289, จำนวนของแหล่งข้อมูลอ้างอิง 218, และความยาวของบทความ 187,746 ตัวอักษร

ลักษณะเด่นแบบสถิติอีกลักษณะหนึ่งที่ถูกใช้ในงานวิจัยนี้คือ จำนวนคำสำคัญที่ปรากฏอยู่ในบทความ โดยคำสำคัญของแต่ละโดเมนได้มาจากขั้นตอนการกำหนดคอนเซ็ปต์ที่ได้กล่าวไว้ก่อนหน้านี้แล้ว เมื่อได้รายการคำสำคัญในโดเมนใด ๆ แล้วจะทำการนับจำนวนคำสำคัญที่ปรากฏในบทความ แล้วเก็บชุดข้อมูลบทความทั้งหมดในรูปแบบเวกเตอร์ของบทความ โดยกำหนดให้บทความแต่ละเรื่องเปรียบเสมือนเวกเตอร์ของคำสำคัญ ขนาดของเวกเตอร์เท่ากับจำนวนของคำสำคัญของโดเมนนั้น ๆ ทั้งนี้เวกเตอร์ที่ประกอบด้วยจำนวนคำสำคัญจะถูกนำไปใช้สำหรับตัวจำแนกแบบเบย์อย่างง่าย

ตารางที่ 4 รายการลักษณะเด่นแบบสถิติ

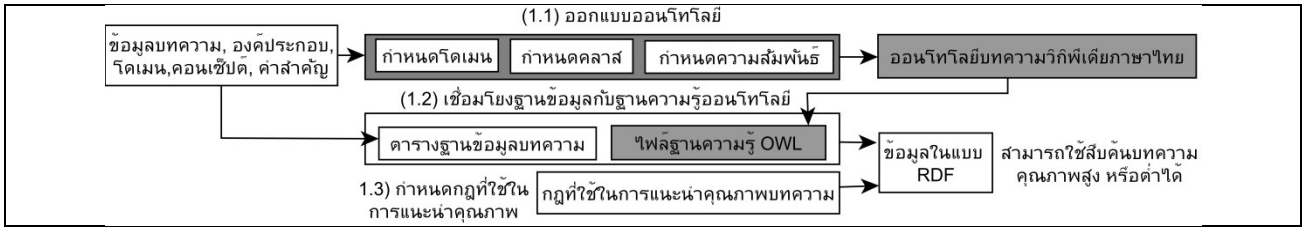
ลักษณะเด่น	คำอธิบาย
1. จำนวนการเชื่อมโยงวิกิ (Wiki Links)	จำนวนการเชื่อมโยงจากหน้าบทความปัจจุบันไปยังหน้าบทความวิกิพีเดียอื่น
2. จำนวนการเชื่อมโยงเว็บเพจ (URL Links)	จำนวนการเชื่อมโยงจากหน้าบทความปัจจุบันไปยังหน้าเว็บเพจอื่น
3. จำนวนเฉลี่ยของการเชื่อมโยงโดเมนเนม	จำนวนเฉลี่ยของการเชื่อมโยงโดเมนจากหน้าบทความปัจจุบันไปยังโดเมนเนมอื่น
4. จำนวนครั้งของการอ้างอิง (References)	จำนวนครั้งของการอ้างอิงไปยังแหล่งข้อมูลอ้างอิงที่บทความปัจจุบันได้อ้างอิงถึง
5. จำนวนของแหล่งข้อมูลอ้างอิง (Unique References)	จำนวนของแหล่งข้อมูลอ้างอิงที่บทความปัจจุบันได้อ้างอิงถึง
6. ความยาวของบทความ	จำนวนของอักขระที่ปรากฏในบทความโดยแทนอักขระด้วยไบต์ (Bytes)

2. การจำแนกคุณภาพบทความ

งานวิจัยนี้นำเสนอวิธีการจำแนกคุณภาพบทความโดยการใช้ลักษณะเด่นแบบคอนเซ็ปต์และแบบสถิติ ลักษณะเด่นแบบคอนเซ็ปต์นำไปใช้สร้างฐานความรู้ออนโทโลยี ลักษณะเด่นแบบสถิติถูกใช้ในวิธีการเรียนรู้ด้วยเครื่องจักร การใช้ลักษณะเด่นทั้งสองแบบจะเป็นประโยชน์ในการตรวจสอบคุณภาพบทความได้ทั้งในแง่ของเนื้อหาและองค์ประกอบอื่นที่จำเป็นของบทความวิกิพีเดีย

2.1 การจำแนกบทความโดยอาศัยฐานความรู้ออนโทโลยี

ฐานความรู้ออนโทโลยีเป็นรูปแบบองค์ความรู้เฉพาะทาง โดยการรวมแนวความคิดสำคัญที่จำเป็นในการอธิบายเป้าหมายของสิ่งหนึ่ง และแสดงความสัมพันธ์ระหว่างแนวความคิดเหล่านั้น ออนโทโลยีในงานวิจัยนี้มีขอบเขตของงานคือต้องการนำออนโทโลยีไปใช้ประโยชน์ในการการจำแนกคุณภาพบทความวิกิพีเดียภาษาไทย การทำงานประกอบด้วย 3 ขั้นตอนหลัก คือ 1. การออกแบบออนโทโลยี 2. การเชื่อมโยงฐานข้อมูลกับออนโทโลยี 3. การกำหนดกฎที่ใช้การแนะนำคุณภาพของบทความ แสดงได้ดังรูปที่ 6



รูปที่ 6 การจำแนกคุณภาพบทความด้วยวิธีออนโทโลยี

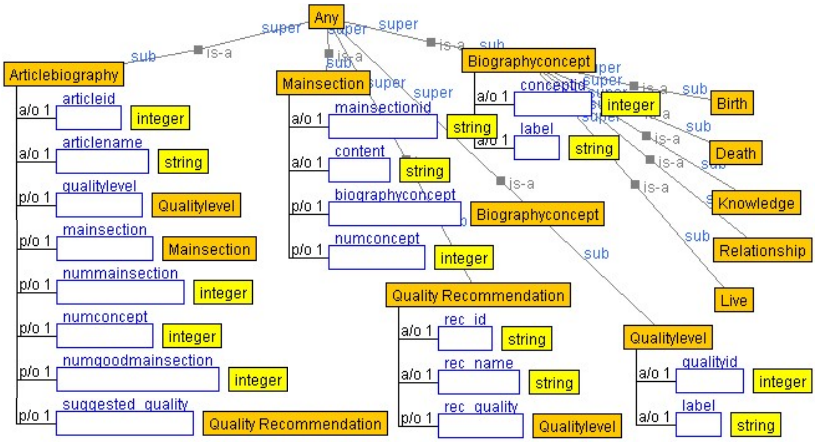
2.1.1 การออกแบบออนโทโลยี

การออกแบบออนโทโลยีในที่นี้ประกอบด้วย 3 ขั้นตอน คือ การกำหนดโดเมน การกำหนดคลาสและลำดับชั้นของคลาส การกำหนดคุณสมบัติและความสัมพันธ์ ดังที่จะได้กล่าวต่อไปนี้

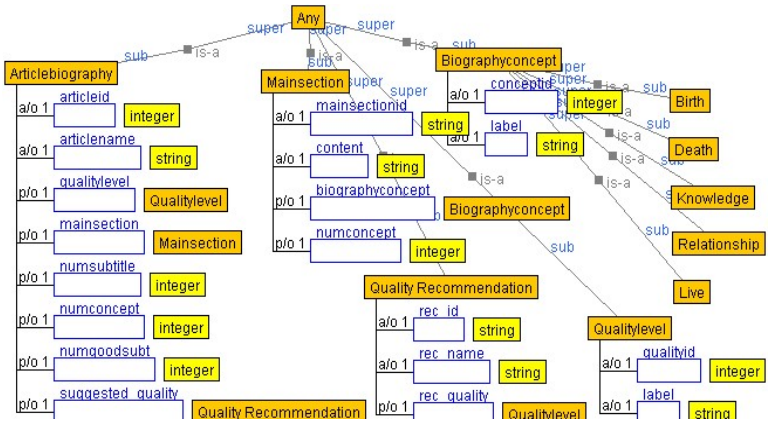
1. การกำหนดโดเมน ในงานวิจัยนี้อาศัยผู้เชี่ยวชาญออกแบบออนโทโลยีในแต่ละโดเมนคือ โดเมนบุคคล โดเมนสถานที่ และโดเมนสัตว์

2. การกำหนดคลาส และลำดับชั้นของคลาส ในงานวิจัยนี้ออนโทโลยีของแต่ละโดเมนมีคลาสหลัก 4 คลาส คือ คลาสบทความ (Article) คลาสเนื้อหาภายใต้หัวข้อหลัก (Mainsection) คลาสคอนเซ็ปต์ (Concept) และคลาสระดับคุณภาพของบทความ (Qualitylevel) ภายใต้คลาสคอนเซ็ปต์ถูกกำหนดให้มีคลาสย่อยเป็นชื่อคอนเซ็ปต์ แสดงดังรูปที่ 7 ออนโทโลยีของโดเมนบุคคล รูปที่ 8 ออนโทโลยีของโดเมนสถานที่ และรูปที่ 9 ออนโทโลยีของโดเมนสัตว์

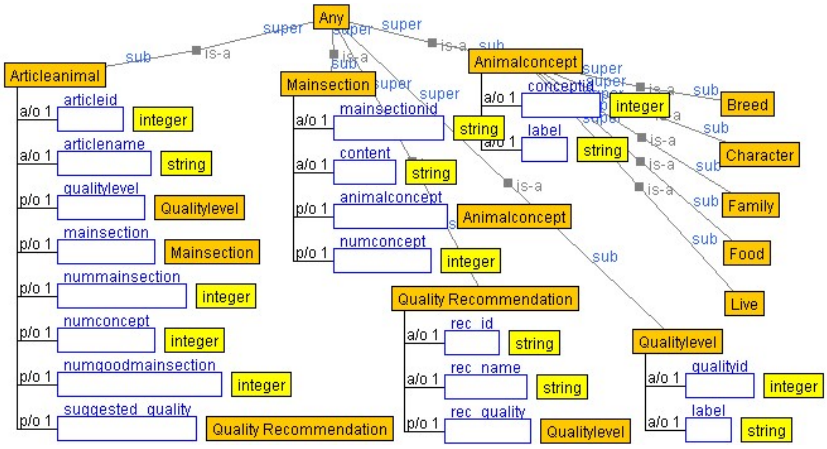
3. การกำหนดคุณสมบัติและความสัมพันธ์ ตามมาตรฐาน OWL ความสัมพันธ์และคุณสมบัติของความสัมพันธ์ประกอบด้วย ความสัมพันธ์แบบจัดเป็น (“is-a”) ความสัมพันธ์แบบการเป็นส่วนหนึ่งของ (“part-of”) และความสัมพันธ์แบบการเป็นคุณลักษณะของ (“attribute-of”)



รูปที่ 7 ออนโทโลยีของโดเมนบุคคล



รูปที่ 8 ออนโทโลยีของโดเมนสถานที่



รูปที่ 9 ออนโทโลยีของโดเมนสัตว์

ความสัมพันธ์แบบจัดเป็น (“is-a”) คือความสัมพันธ์แบบที่มีคุณสมบัติของการถ่ายทอดคุณสมบัติของคลาสหนึ่งไปยังอีกคลาสหนึ่ง เช่น รูปที่ 8 ออนโทโลยีของโดเมนสถานที่ แสดงความสัมพันธ์ของคลาสตำแหน่งที่ตั้ง (Sitestructure) “is-a” คลาสคอนเซ็ปต์สถานที่ (Placeconcept) ซึ่งอธิบายได้ว่าตำแหน่งที่ตั้งเป็นชื่อคอนเซ็ปต์หนึ่งของคอนเซ็ปต์สถานที่

ความสัมพันธ์แบบการเป็นส่วนหนึ่งของ (“part-of”) คือความสัมพันธ์ที่หมายถึงการเป็นส่วนประกอบ เช่นรูปที่ 7 คลาสระดับคุณภาพบทความ (Qualitylevel) “part-of” คลาส

บทความ (Articlebiography) ซึ่งอธิบายได้ว่าระดับคุณภาพบทความเป็นส่วนหนึ่งของบทความ

ความสัมพันธ์แบบการเป็นคุณลักษณะของ (attribute-of) เป็นความสัมพันธ์ของข้อมูลที่มีคุณสมบัติชนิดใดชนิดหนึ่ง ตัวอย่างคุณสมบัติของข้อมูลเช่น ตัวเลข ตัวอักษร วันที่ เป็นต้น ตัวอย่างความสัมพันธ์เช่นรูปที่ 7 โหนด ชื่อบทความ (articlename) attribute-of คลาสบทความ (Articlebio-graphy) ซึ่งอธิบายได้ว่า ชื่อบทความเป็นคุณลักษณะของบทความ ซึ่งชื่อบทความเป็นข้อมูลชนิดสายอักขระ (string)

งานวิจัยนี้ใช้ Hozo (Kozaki et al., 2002) เป็นเครื่องมือในการพัฒนาออนโทลียีบทความวิกิพีเดียภาษาไทยและจัดเก็บเป็นไฟล์ฐานความรู้ตามมาตรฐาน OWL (Ontology Web Language) ออนโทโลยีของแต่ละโดเมนแสดงได้ดังรูปที่ 7 แสดงออนโทโลยีของโดเมนบุคคล รูปที่ 8 แสดงออนโทโลยีของโดเมนสถานที่ และรูปที่ 9 แสดงออนโทโลยีของโดเมนสัตว์ สำหรับคลาสแนะนำคุณภาพ (Quality Recommendation) เป็นคลาสที่สร้างเพิ่มเติมลงไปในออนโทโลยีบทความวิกิพีเดียภาษาไทยในภายหลัง เพื่ออำนวยความสะดวกในแสดงผลการสืบค้นข้อมูลหลังจากที่ได้มีการกำหนดกฎสำหรับการสืบค้นในขั้นตอนที่ (1.3) แล้วคลาสดังกล่าวประกอบด้วยหมายเลขการแนะนำคุณภาพ (rec_id) และชื่อการแนะนำคุณภาพ (rec_name) ซึ่งมีความสัมพันธ์แบบการเป็นคุณลักษณะของ (attributes-of) และคลาสดังกล่าวยังประกอบด้วยคุณภาพที่แนะนำ (rec_quality) ซึ่งมีความสัมพันธ์แบบการเป็นส่วนหนึ่งของ (part-of)

2.1.2 การเชื่อมโยงฐานข้อมูลกับฐานความรู้ออนโทโลยี

เพื่อให้สามารถนำออนโทโลยีวิกิพีเดียภาษาไทยที่สร้างขึ้นไปใช้งานในโปรแกรมประยุกต์ต่าง ๆ ตามแบบเทคโนโลยีเว็บเชิงความหมาย (Semantic Web Technology) ได้อย่างสะดวกขึ้น จึงทำการเชื่อมโยงความสัมพันธ์ระหว่างข้อมูล (Data Mapping) ในฐานข้อมูลกับฐานความรู้ออนโทโลยีที่ถูกจัดเก็บตามมาตรฐาน OWL ในงานวิจัยนี้อาศัยระบบจัดการโปรแกรมประยุกต์ฐานความรู้ออนโทโลยี หรือ OAM (Ontology Application Management) (Buranarach et al., 2016) เป็นเครื่องมือในการเชื่อมโยงข้อมูลเข้ากับโครงสร้างออนโทโลยี เมื่อทำการเชื่อมโยงฐานข้อมูลบทความวิกิพีเดียภาษาไทยกับออนโทโลยีบทความวิกิพีเดียภาษาไทย แล้วระบบจะสร้างข้อมูลขึ้นมาในรูปแบบ RDF (Resource Description Framework) ซึ่งเป็นข้อมูลที่สามารถนำไปใช้ประโยชน์ต่อไปได้ในงานด้านเทคโนโลยีเว็บเชิงความหมาย

ฐานข้อมูลบทความวิกิพีเดียภาษาไทยและออนโทโลยีบทความวิกิพีเดียภาษาไทยถูกสร้างเตรียมไว้แล้วในขั้นตอนก่อนหน้านี้ โดยแสดงโครงสร้างตารางฐานข้อมูลในรูปที่ 5 ซึ่งโครงสร้างตารางในแต่ละโดเมน ถูกออกแบบให้มีความสอดคล้องกันกับคลาสหลัก คลาสย่อย และคุณสมบัติของออนโทโลยีแล้ว เพื่อให้เกิดความง่ายต่อการเชื่อมโยงฐานข้อมูลกับออนโทโลยีโดยอาศัย OAM ทั้งนี้ การเชื่อมโยงข้อมูลประกอบด้วย 3 ส่วน คือ การกำหนดความสัมพันธ์ของคลาสกับตาราง การกำหนดความสัมพันธ์ของคุณสมบัติกับคอลัมน์ และการกำหนดค่าการแปลงคำศัพท์

1. การกำหนดความสัมพันธ์ของคลาสกับตาราง ตัวอย่างเช่นรูปที่ 7 ในโดเมนบุคคล กำหนดคลาสบทความ (Articlebiography) ให้สัมพันธ์กับตารางบทความในรูปที่ 5 กำหนดคลาสหัวข้อหลัก (Mainsection) ให้สัมพันธ์กับตารางหัวข้อหลักในรูปที่ 5 เป็นต้น

2. การกำหนดความสัมพันธ์ของคุณสมบัติกับคอลัมน์ ตัวอย่างเช่นรูปที่ 7 กำหนดความสัมพันธ์ของหมายเลขคุณภาพ (qualityid) กับคอลัมน์หมายเลขคุณภาพของบทความในรูปที่ 5 กำหนดความสัมพันธ์ของลาเบล (label) ในรูปที่ 7 กับชื่อคุณภาพของบทความในรูปที่ 5 เป็นต้น

3. การกำหนดค่าการแปลงคำศัพท์ (Vocabulary Mapping) ตัวอย่างเช่นรูปที่ 7 ภายใต้คลาสข้อมูลการเกิด (Birth) ซึ่งอยู่ในคลาสคอนเซ็ปต์ของโดเมนบุคคล (Biography-concept) สามารถกำหนดค่าการแปลงคำศัพท์สำหรับรายการคำได้ โดยนำคำว่า เกิดวันที่, สถานที่เกิด, บ้านเกิด จากรายการคำในรูปที่ 5 ส่งไปแปลงให้เป็นค่าของคลาสข้อมูลการเกิดได้

เมื่อออนโทโลยีเชื่อมโยงกับฐานข้อมูลแล้วจะถูกเก็บไว้ในรูปแบบมาตรฐาน RDF ซึ่งจะสามารถสืบค้นข้อมูลเชิงความหมายได้ โดยสามารถสืบค้นบทความได้ตามคลาส ตามคุณสมบัติของคลาส และตามความสัมพันธ์ของคลาส ซึ่งจะตั้งค่าการสืบค้นผ่าน OAM ได้ นอกจากนี้ระบบดังกล่าวยังสนับสนุนกลไกการอนุมานผ่านกฎ (Rule-Based Inference) เพื่อเพิ่ม

ความสามารถในการแนะนำข้อมูลด้วย ดังที่จะได้กล่าวในหัวข้อต่อไป

2.1.3 การกำหนดกฎที่ใช้การแนะนำคุณภาพของบทความ

การเพิ่มความถูกต้องในการจำแนกคุณภาพบทความของงานวิจัยนี้อาศัยกฎที่สอดคล้องกับสมมุติฐานประการหนึ่งของงานวิจัยนี้คือ บทความคุณภาพดีควรมีเนื้อหาที่ครอบคลุมคอนเซ็ปต์ส่วนใหญ่ของโดเมนนั้น และเนื้อหาภายใต้หัวข้อหลักแต่ละหัวข้อของบทความควรมีเพียงหนึ่งคอนเซ็ปต์ ซึ่งในแต่ละคอนเซ็ปต์ประกอบด้วยคำสำคัญที่เป็นตัวแทนของคอนเซ็ปต์ได้

การกำหนดกฎสำหรับการทดลองอาศัยอัตราส่วนของคำสำคัญของคอนเซ็ปต์ในหัวข้อหลัก เพื่อปรับปรุงการพิจารณาว่าหัวข้อหลักมีคอนเซ็ปต์อยู่หรือไม่ เงื่อนไขที่ถูกกำหนดขึ้นในการทดลองนี้ประกอบด้วย 3 เงื่อนไข แสดงในรูปที่ 10 โดยกำหนดให้ i แทนเนื้อหาภายใต้หัวข้อหลักของบทความ A กำหนดให้ r_c คือ จำนวนเฉลี่ยของคำในคอนเซ็ปต์ c ใด ๆ กับจำนวนคำที่ปรากฏในเนื้อหาภายใต้หัวข้อหลักนั้น กำหนดให้ AVGi(r_c) คือค่าเฉลี่ยของ r_c กับจำนวนคอนเซ็ปต์ของโดเมนนั้น

The first condition:

The concept found in A_i = $\begin{cases} 1, if(r_c \geq AVG_i(r_c)). \\ 0, otherwise. \end{cases}$ As a concept vector of A_i .

The second condition:

The quality of A_i = $\begin{cases} good, if(the\ sum\ of\ the\ concept\ vector\ of\ A_i \leq 2). \\ not\ good, otherwise. \end{cases}$

The third condition: (n *is the total number of* A_i)

The quality of *artice A*= $\begin{cases} low, if(n \leq 2). \\ high, if(n > 2) \wedge (the\ number\ of\ A_i\ as\ the\ good\ section \geq 60\%\ of\ n) \\ low, otherwise. \end{cases}$

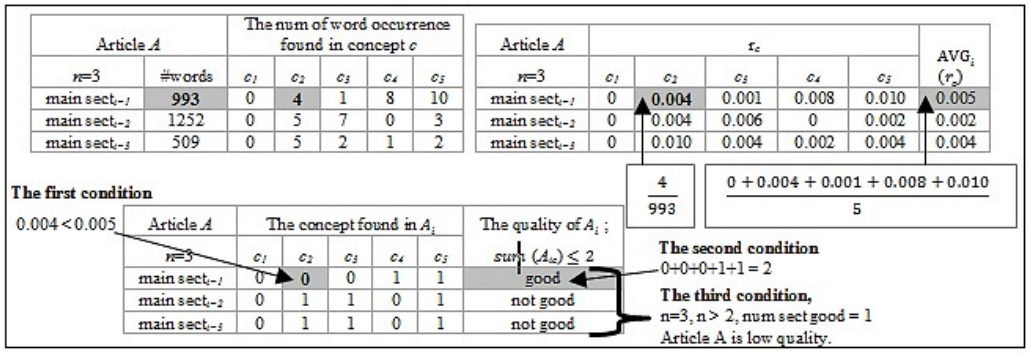
รูปที่ 10 วิธีการใช้กฎจากออนโทโลยี

เงื่อนไขที่ 1 กำหนดมาเพื่อลดความสำคัญของคอนเซ็ปต์ที่มีคำปรากฏจำนวนน้อย กล่าวคือถ้า $r_c \geq$ AVGi(r_c) คอนเซ็ปต์ของเนื้อหาภายใต้หัวข้อหลักส่วนนั้นจะเป็น 0 และจะเป็น 1 หากมีคำสำคัญมากพอ ทั้งนี้ จะได้คอนเซ็ปต์เว็กเตอร์ของเนื้อหาภายใต้หัวข้อหลัก

เงื่อนไขที่ 2 พิจารณาจำนวนคอนเซ็ปต์ของเนื้อหาภายใต้หัวข้อหลัก ซึ่งสอดคล้องกับสมมุติฐานของงานวิจัยคือควรมีคอนเซ็ปต์เดียว โดยคำนวณจากผลรวมของคอนเซ็ปต์เว็กเตอร์ของเนื้อหาภายใต้หัวข้อหลักนั้น หากจำนวนคอนเซ็ปต์$\leq$ 2 ถือว่าเนื้อหาภายใต้หัวข้อหลักนั้นเป็นเนื้อหาที่ดี

เงื่อนไขที่ 3 พิจารณาจำนวนหัวข้อหลักซึ่งมีเนื้อหาที่ดี กล่าวคือ หากบทความมีจำนวนหัวข้อหลักน้อยกว่าหรือเท่ากับ 2 หัวข้อ ถือได้ว่าเป็นบทความที่คุณภาพต่ำ หากจำนวนหัวข้อมีมากกว่า 2 จะพิจารณาว่าหากจำนวนหัวข้อหลักที่มีเนื้อหาดีมีจำนวนมากกว่าร้อยละ 60 ของจำนวนหัวข้อหลักทั้งหมดในบทความ จึงจะเป็นบทความที่มีคุณภาพสูง หากไม่เป็นเช่นนั้นถือได้ว่าเป็นบทความที่มีคุณภาพต่ำ

จากเงื่อนไขที่กำหนด แสดงตัวอย่างการคำนวณดังรูปที่ 11 โดยกำหนดให้บทความ A เป็นบทความในโดเมนบุคคล มีคอนเซ็ปต์ทั้งหมด 5 คอนเซ็ปต์ (c_1,c_2,c_3,c_4,c_5) ประกอบด้วย 3 หัวข้อหลัก (main $sect_i$) (n=3) เนื้อหาในหัวข้อหลักที่ 1 มีคำทั้งหมดจำนวน 993 คำ ในหัวข้อหลักที่ 2 มีคำจำนวนทั้งหมด 1,252 คำ ในหัวข้อหลักที่ 3 มีคำจำนวนทั้งหมด 509 คำเงื่อนไขที่สร้างขึ้นจะถูกนำเข้าไปในระบบจัดการแนะนำข้อมูล (Recommendation Management System) ของ OAM ซึ่งสนับสนุนกลไกการอนุมานผ่านกฎ (Rule-based Inference) โดยทำงานกับข้อมูล RDF และสร้างข้อมูลที่สอดคล้องกับกฎเพื่อให้ได้ผลลัพธ์ตามที่ต้องการที่สามารถนำไปใช้ในการสืบค้นข้อมูลต่อไปได้



รูปที่ 11 ตัวอย่างการคำนวณตามกฎที่ใช้ในการแนะนำคุณภาพบทความ

2.2 การจำแนกบทความโดยวิธีการเรียนรู้ด้วยเครื่องจักร

งานวิจัยนี้นำวิธีการเรียนรู้ด้วยเครื่องจักรมาเพิ่มประสิทธิภาพให้กับการจำแนกบทความโดยอาศัยฐานความรู้ออนโทโลยี วิธีการเรียนรู้ด้วยเครื่องจักรที่ใช้ในงานวิจัยนี้คือ ตัวจำแนกต้นไม้ตัดสินใจ และตัวจำแนกแบบเบย์อย่างง่าย

2.2.1 วิธีการที่งานวิจัยนี้นำเสนอ

งานวิจัยนี้นำเสนอการจำแนกคุณภาพบทความโดยใช้ฐานความรู้ออนโทโลยีร่วมกับการใช้ตัวจำแนกต้นไม้ตัดสินใจ ระบบการจำแนกประกอบด้วย 2 ขั้นตอนคือ เริ่มจากการคัดกรองคุณภาพโดยอาศัยฐานความรู้ออนโทโลยีที่มีการกำหนดกฎให้กับระบบการแนะนำคุณภาพบทความ เมื่อได้บทความที่ระบบแนะนำว่าเป็นบทความคุณภาพสูงแล้วจะทำงานในขั้นตอนที่สองคือ นำชุดบทความคุณภาพสูงไปจำแนกคุณภาพอีกครั้งหนึ่งด้วยตัวจำแนกต้นไม้ตัดสินใจ (แสดงในรูปที่ 1 โดยวิธีเรียนรู้ด้วยเครื่องจักรในรูปนั้นคือตัวจำแนกต้นไม้ตัดสินใจ) ผู้วิจัยได้นำตัวอย่างสำหรับการเรียนรู้ทั้งหมดมาเรียนรู้ด้วยตัวจำแนกต้นไม้ตัดสินใจ พบว่าได้ลักษณะเด่นแสดงดังตารางที่ 4 โดยบทความในแต่ละโดเมนมีลักษณะเด่นที่แตกต่างกัน กล่าวคือ จำนวนการเชื่อมโยงเว็บเพจ จำนวนเฉลี่ยของการเชื่อมโยงโดเมน จำนวนครั้งของการอ้างอิง และความยาวของบทความ เป็นลักษณะเด่นที่ปรากฏในโดเมนบุคคล แต่จำนวนการเชื่อมโยงวิกิเป็นลักษณะเด่นที่ปรากฏในโดเมนสถานที่เพียงโดเมนเดียว และพบว่าจำนวนของแหล่งข้อมูลอ้างอิงที่เป็นลักษณะเด่นที่ปรากฏในทั้งสามโดเมน (Soonthornphisaj and Paengporn, 2017) กฎที่ได้จากต้นไม้ตัดสินใจถูกนำไปใช้ในขั้นตอนหลังจากผ่านการคัดกรองด้วยออนโทโลยีแล้ว

2.2.2 การจำแนกบทความโดยอาศัยฐานความรู้ออนโทโลยีร่วมกับตัวจำแนกแบบเบย์อย่างง่าย

เป็นการจำแนกคุณภาพบทความโดยใช้ 2 ขั้นตอน ทำนองเดียวกับวิธีการที่งานวิจัยนี้นำเสนอ แตกต่างที่ทดลองร่วมกับตัวจำแนกแบบเบย์อย่างง่ายแทนการใช้ต้นไม้ตัดสินใจ ชุดข้อมูลที่ใช้สำหรับตัวจำแนกแบบเบย์อย่างง่ายประกอบด้วยเวกเตอร์ของคำสำคัญที่ปรากฏอยู่ในบทความ จำนวนคำสำคัญที่ปรากฏในบทความในแต่ละประเภทของคุณภาพบทความจะถูกใช้ในการคำนวณหาความน่าจะเป็นที่บทความจะมีคุณภาพสูงหรือต่ำได้

3. การประเมินผลการจำแนกคุณภาพบทความ

เพื่อประเมินประสิทธิภาพของการทดลอง ในงานวิจัยนี้ใช้การวัดค่าความแมนยำ (Precision) การวัดค่าความระลึก (Recall) และการวัดค่าประสิทธิภาพ (*F*-Measure หรือ F_1 *score*) แสดงในสมการที่ (4) สมการที่ (5) และสมการที่ (6) ทั้งนี้ได้ทำการแบ่งชุดข้อมูลออกเป็น 10 ส่วน ด้วยวิธีการทดสอบแบบไขว้ข้าม *k* กลุ่ม (*k*-Fold Cross Validation) สำหรับตัวอย่างเรียนรู้ (Training Data) และตัวอย่างทดสอบ (Test Data)

$$Precision = \frac{No.of\ True\ Positive}{No.of\ True\ Positive + No.of\ False\ Positive} \quad (4)$$

$$Recall = \frac{No.of\ True\ Positive}{No.of\ True\ Positive + No.of\ False\ Negative} \quad (5)$$

$$F - Measure = \frac{2x\ Precision x\ Recall}{Precision + Recall} \quad (6)$$

ผลการวิจัย

จากผลการทดลองพบว่าจำนวนบทความที่ถูกจำแนกเป็นบทความคุณภาพสูงด้วยกฎจากออนโทโลยี ในโดเมนบุคคลเท่ากับ 1,895 ในโดเมนสัตว์เท่ากับ 41 และในโดเมนสถานที่เท่ากับ 1,545 บทความ ซึ่งบทความเหล่านี้จะถูกตัวจำแนกต้นไม้ตัดสินใจที่ใช้ลักษณะเด่นเกี่ยวกับการอ้างอิงทำการจำแนกคุณภาพอีกครั้งหนึ่ง พบว่าจำนวนบทความคุณภาพสูงซึ่งได้จากวิธีการที่งานวิจัยนี้นำเสนอ ในโดเมนบุคคลเท่ากับ 64 ในโดเมนสัตว์เท่ากับ 9 และในโดเมนสถานที่เท่ากับ 34 บทความ แสดงให้เห็นว่าวิธีการที่งานวิจัยนี้นำเสนอ ให้ค่าประสิทธิภาพสูงที่สุดเมื่อเทียบกับวิธีการอื่น ๆ ซึ่งแสดงผลการทดลองตามตารางที่ 5 ค่าความแม่นยำ (Precision) ในโดเมนบุคคลเท่ากับ 0.859 ในโดเมนสัตว์เท่ากับ 0.890 และในโดเมนสถานที่เท่ากับ 0.650 ซึ่งสูงกว่าวิธีออนโทโลยีร่วมกับตัวจำแนกแบบเบย์อย่างง่าย โดยให้ค่าความแม่นยำในโดเมนบุคคลเท่ากับ 0.239 ในโดเมนสัตว์เท่ากับ 0.750 และในโดเมนสถานที่เท่ากับ 0.170 วิธีการออนโทโลยีได้ให้ค่าความแม่นยำในโดเมนบุคคลเท่ากับ 0.033 ในโดเมนสัตว์เท่ากับ 0.195 และในโดเมนสถานที่เท่ากับ 0.017 และวิธีจำแนกคุณภาพบทความด้วยตัวจำแนกแบบเบย์อย่างง่ายให้ค่าความแม่นยำในโดเมนบุคคลเท่ากับ 0.105 ในโดเมนสัตว์เท่ากับ 0.391 และในโดเมนสถานที่เท่ากับ 0.076 สำหรับผลการจำแนกคุณภาพบทความด้วยตัวจำแนกต้นไม้ตัดสินใจแสดงไว้ในการทดลองนี้ (Soonthornphisaj and Paengporn, 2017) ซึ่งพบว่าสามารถจำแนกคุณภาพบทความตามลักษณะเด่นที่เกี่ยวกับการอ้างอิงได้ค่าความแม่นยำสูง อย่างไรก็ตาม จำเป็นต้องพิจารณาเนื้อหาของบทความด้วย จึงนำกฎจากออนโทโลยีมาคัดกรองบทความที่มีเนื้อหาครอบคลุมทุกคอนเซ็ปต์ก่อน แล้วจึงพิจารณาองค์ประกอบเกี่ยวกับการอ้างอิงต่อไป ทั้งนี้เพื่อให้ได้บทความคุณภาพสูงตามเกณฑ์ที่วิกิพีเดียกำหนด

ตารางที่ 5 ผลการทดลองการจำแนกคุณภาพบทความวิกิพีเดียภาษาไทย

	วิธีการที่งานวิจัยนี้นำเสนอ		**ออนโทโลยีร่วมกับตัวจำแนกแบบเบย์อย่างงาย**		**ออนโทโลยี**		**ตัวจำแนกแบบเบย์อย่างง่าย**	
คุณภาพ	สูง	ต่ำ	สูง	ต่ำ	สูง	ต่ำ	สูง	ต่ำ
				โดเมนบุคคล				
จำนวนบทความ	64	8,598	205	8,457	1,895	6,767	617	8,045
Precision	**0.859**	0.999	0.293	1.000	0.033	1.000	0.105	1.000
Recall	**0.833**	0.999	0.909	0.980	0.955	0.787	0.985	0.936
F-measure	**0.846**	0.999	0.443	0.990	0.064	0.881	0.190	0.967
				โดเมนสัตว์				
จำนวนบทความ	9	255	8	256	41	223	23	241
Precision	**0.890**	1.000	0.750	0.990	0.195	0.996	0.391	1.000
Recall	**0.890**	1.000	0.670	0.990	0.889	0.871	1.000	0.945
F-measure	**0.890**	1.000	0.710	0.990	0.320	0.929	0.563	0.972
				โดเมนสถานที่				
จำนวนบทความ	34	6,036	138	5,932	1,545	4,525	353	5,717
Precision	**0.650**	1.000	0.170	1.000	0.017	1.000	0.076	1.000
Recall	**0.810**	1.000	0.890	0.980	1.000	0.749	1.000	0.946
F-measure	**0.720**	1.000	0.290	0.990	0.034	0.856	0.142	0.972

วิจารณ์ผลการวิจัย

ผู้วิจัยพบว่าการใช้ตัวจำแนกแบบเบย์อย่างง่ายไม่สามารถจำแนกบทความในเชิงคุณภาพได้ ถึงแม้ว่าอัลกอริทึมตัวจำแนกแบบเบย์อย่างง่ายจะมีประสิทธิภาพสูงในการจำแนกประเภทเอกสารในโดเมนที่ต่างกัน แต่ลักษณะของข้อมูลในงานวิจัยนี้เป็นการจำแนกบนความบนโดเมนเดียวกันออกเป็นสองประเภทคือคุณภาพสูงและคุณภาพต่ำ ซึ่งการใช้ตัวจำแนกแบบเบย์อย่างง่ายเพียงอย่างเดียวให้ค่าประสิทธิภาพในการจำแนก

ด้อยกว่าการทำงานร่วมกับวิธีการใช้ออนโทโลยี พบว่าในการจำแนกบทความคุณภาพสูง ตัวจำแนกแบบเบย์อย่างง่ายให้ความแม่นยำน้อยกว่า 2 เท่า และพบว่าการใช้ออนโทโลยีร่วมกับตัวจำแนกแบบเบย์อย่างง่ายให้ค่าประสิทธิภาพสูงกว่าการใช้ตัวจำแนกแบบเบย์อย่างง่ายเพียงอย่างเดียวถึง 8 เท่าในโดเมนคน 2 เท่าในโดเมนสัตว์ และมากกว่าถึง 9 เท่าในโดเมนสถานที่

สำหรับการใช้กฎจากออนโทโลยี ผู้วิจัยพบว่าค่าความแม่นยำของบทความคุณภาพต่ำเป็นค่าที่สูงมาก (1, 0.996, 1 สำหรับโดเมนบุคล โดเมนสัตว์ และโดเมนสถานที่ ตามลำดับ) และให้ค่าความถูกต้อง (Recall) ของบทความ คุณภาพสูงด้วยค่าที่สูง (0.955, 0.889, 1.000 สำหรับโดเมนบุคคล โดเมนสัตว์ และโดเมนสถานที่ ตามลำดับ) แสดงว่าวิธีการออนโทโลยีเพียงอย่างเดียวสามารถเลือกบทความที่มีคุณภาพสูงได้เกือบทั้งหมดจากบทความที่มีทั้งหมด ในขณะที่เมื่อบทความมีคุณภาพต่ำก็สามารถให้ค่าคุณภาพต่ำกับบทความที่มีคุณภาพต่ำได้อย่างถูกต้อง ดังนั้น บทความที่ผ่านการคัดกรองด้วยกฎจากออนโทโลยีแล้วจะมีบทความที่มีคุณภาพสูงและต่ำปะปนกันไป โดยมีสัดส่วนของบทความคุณภาพสูงน้อยกว่าบทความคุณภาพต่ำเป็นจำนวนมาก (สังเกตได้จากค่าความแม่นยำของบทความคุณภาพสูง 0.033, 0.195, 0.017 สำหรับโดเมนบุคล โดเมนสัตว์ และโดเมนสถานที่) ถึงแม้การใช้กฎจากออนโทโลยีร่วมกับวิธีการเรียนรู้ด้วยเครื่องจักรจะสามารถเพิ่ม ประสิทธิภาพให้กับการจำแนกบทความได้ แต่อย่างไรก็ตาม จำนวนบทความวิกิพีเดียภาษาไทยมีปัญหาความไม่สมดุล ดังนั้น แนวทางการวิจัยต่อไปอาจจะพิจารณาใช้เทคนิคการแก้ปัญหาตัวอย่างข้อมูลที่ไม่สมดุลเพื่อให้ผลการจำแนกโดยรวมดีขึ้น

กิตติกรรมประกาศ

งานนี้วิจัยนี้ได้รับการสนับสนุนจากสถาบันวิจัยและพัฒนามหาวิทยาลัยเกษตรศาสตร์ (KURDI: Kasetsart University Research and Development) และขอขอบคุณ ดร.มารุต บูรณรัช จากห้องปฏิบัติการวิจัยเทคโนโลยีภาษาธรรมชาติและความหมาย ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติที่ให้การสนับสนุนด้านเทคนิคในการพัฒนาออนโทโลยี

เอกสารอ้างอิง

วิกิพีเดีย. (2558). บทความคัดสรรคืออะไร. แหล่งข้อมูล: https://th.wikipedia.org/wiki/วิกิพีเดีย:บทความคัดสรรคืออะไร. ค้นเมื่อวันที่ 28 พฤษภาคม 2560.

วิกิพีเดีย. (2556). บทความคุณภาพคืออะไร. แหล่งข้อมูล: https://th.wikipedia.org/wiki/วิกิพีเดีย:บทความบทความคุณภาพคืออะไร. ค้นเมื่อวันที่ 28 พฤษภาคม 2560.

วิกิพีเดีย. (2560). สถิติ. แหล่งข้อมูล: https://th.wikipedia.org/wiki/พิเศษ:สถิติ. ค้นเมื่อวันที่ วันที่ 6 มิถุนายน 2560.

วิกิพีเดีย. (2559). ความแตกต่างระหว่างบทความคัดสรรและบทความคุณภาพ. แหล่งข้อมูล: https://th.wikipedia.org/wiki/วิกิพีเดีย:ความแตกต่างระหว่างบทความคัดสรรและบทความคุณภาพ. ค้นเมื่อวันที่ 28 พฤษภาคม 2560.

Betancourt, G.G., Segnine, A., Trabuco, C., Rezgui, A. and Jullien, N. (2016). Mining team characteristics to predict Wikipedia article quality. In: Proceedings of the 12th International Symposium on Open Collaboration, ACM, Berlin, Germany. 1-9.

Buranarach, M., Supnithi, T., Thein, Y.M., Ruangrajitpakorn, T., Rattanasawad, T., Wongpatikaseree, K., Lim, A.O., Tan, Y. and Assawamakin, A. (2016). OAM: An Ontology Application Management Framework for Simplifying Ontology-Based Semantic Web Application Development. International Journal of Software Engineering and Knowledge Engineering. (26): 115-146.

Calzada, G.D.l. and Dekhtyar, A. (2010). On measuring the quality of Wikipedia articles. In: Proceedings of the 4th workshop on Information credibility, ACM, Raleigh, North Carolina, USA. 11-18.

Chevalier, F., Huot, S. and Fekete, J.D. (2010). WikipediaViz: Conveying article quality for casual Wikipedia readers. In: IEEE Pacific Visualization Symposium. 49-56.

Dalip, D.H., Santos, R.L., Renn, D., Oliveira, Val, Amaral, r.F., Andr, M., Gon, alves, Prates, R.O., Minardi, R.C.M. and Almeida, J.M.d. (2011). GreenWiki: a tool to support users' assessment of the quality of Wikipedia articles. In: 11th annual international ACM/IEEE joint conference on Digital libraries, ACM, Ottawa, Ontario, Canada. 469-470.

Dang, Q.V. and Ignat, C.-L. (2016). Quality Assessment of Wikipedia Articles without Feature Engineering. In: 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, ACM, Newark, New Jersey, USA. 27-30.

Harpalani, M., Phumprao, T., Bassi, M., Hart, M. and Johnson, R. (2010). Wiki Vandalysis - Wikipedia Vandalism Analysis. In: Lab Report for PAN at CLEF.

Hartig, O. (2013). SQUIN: a traversal based query execution system for the web of linked data. In: ACM SIGMOD International Conference on Management of Data, ACM, New York, USA. 1081-1084.

Jain, A. K. and Li, Y. H. (1998). Classification of Text Documents. The Computer Journal. (41):537-546.

Javanmardi, S. (2011). Measuring Content Quality in User Generated Content Systems: a Machine Learning Approach. Doctor of philosophy in Information and Computer Science, University of California. Irvine: 175.

Kozaki, K., Kitamura, Y., Ikeda, M. and Mizoguchi, R. (2002). Hozo: An Environment for Building/Using Ontologies Based on a Fundamental Consideration of “Role” and “Relationship”. In: 13th International Conference of Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, Springer Berlin Heidelberg, Berlin, Heidelberg. 213-218.

Li X., Tang J., Wang T., Luo Z. and de Rijke M. (2015). Automatically Assessing Wikipedia Article Quality by Exploiting Article–Editor Networks. In: 37th European Conference on IR Research, Springer International Publishing, Cham. 574-580.

Nora I. Al-Rajebah, Hend S. Al-Khalifa and AbdulMalik S. Al-Salman. (2011). Exploiting Arabic Wikipedia for automatic ontology generation: A proposed approach. In: International Conference on Semantic Technology and Information Retrieval.

Saengthongpattana, K., Soonthornphisaj, N. (2014). Assessing the Quality of Thai Wikipedia Articles Using Concept and Statistical Features. In: New Perspectives in Information Systems and Technologies, Volume 1, Springer International Publishing, Cham. 513-523.

Soonthornphisaj, N., and Paengporn, P. (2017). Thai Wikipedia Article Quality Filtering Algorithm, Lecture Notes in Engineering and Computer Science. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong. 299-305.

Pan, J.Z., Thomas, E. and Sleeman, D. (2006). Ontosearch: Searching and querying web ontologies. In: Proceedings of the IADIS international conference. 211–218.

Sciascio, C.d., Strohmaier, D., Errecalde, M. and Veas, E. (2017). WikiLyzer: Interactive Information Quality Assessment in Wikipedia. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces, ACM, Limassol, Cyprus. 377-388.

Suzuki, Y. and Nakamura, S. (2016). Assessing the Quality of Wikipedia Editors through Crowdsourcing, In: Proceedings of the 25th International Conference Companion on World Wide Web, Canada. 1001-1006.

Tamagawa, S., Sakurai, S., Tejima, T., Morita, T., Izumi, N. and Yamaguchi, T. (2010). Learning a Large Scale of Ontology from Japanese Wikipedia. In: International Conference on Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM. 279-286.

Tzekou, P., Stamou, S., Kirtsis, N. and Zotos, N. (2011). Quality Assessment of Wikipedia External Links. In: the 7th International Conference on Web Information Systems and Technologies, Noordwijkerhout, Netherlands. 248-254.

Vijay V. Raghavan and S. K. M. Wong. (1986). A critical analysis of vector space model for information retrieval. Journal of the American Society for Information Science. 37(5): 279-87.

Wikimedia Foundation. (2017). Compare criteria Good and Featured article. Available: https://en.wikipedia.org/wiki/Wikipedia:Compare_criteria_Good_v._Featured_article. Retrieved 26 June 2017

Wikipedia. (2017). WikiProject assessment. Available: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_assessment. Retrieved 26 June 2017.

❑❑❑❑❑