การจำแนกความน่าเชื่อถือของเนื้อหาในเว็บไซต์ภาษาไทย ด้านมะเร็งโดยใช้ CancerDic+

Classification of Reliable Content on Cancer Thai Website using CancerDic+

สุภาพร เกิดกิจ[1] องอาจ อุ่นอนันต์[2] และ พยุง มีสัจ[3]

[1,3]คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

[2]คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ

ABSTRACT – Nowadays, there are a lot of informative websites about cancer which enables users to access data easily. However, it is difficult to determine whether these websites are reliable. The objective of this research was to develop a method which could classify the credibility of a cancer website and then determine if the information was reliable. To achieve the above goal, this research applied CancerDic+ to extract words. Technical terms referring to cancer were added to a database, and then text mining was applied to classify inputted data. The values of accuracy, precision and recall derived from the word extraction by Lexto, SWATH and CancerDic+ were then compared. The results showed that text mining for word extraction by CancerDic+ yielded the best result (Accuracy = 0.844, Precision = 0.838, Recall =0.845). In conclusion, this classification method successfully gained high accuracy and can be used in other fields effectively.

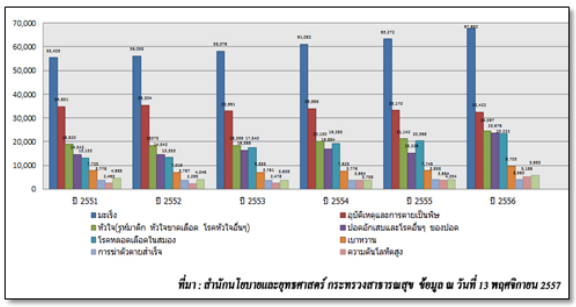
KEYWORDS – Text Mining; Data Classification; Word Extraction; Website Credibility

บทคัดย่อ – ปัจจุบันจำนวนเว็บไซต์ที่ให้ความรู้ด้านมะเร็งมีอยู่เป็นจำนวนมาก ทำให้ผู้ใช้งานเข้าถึงข้อมูลได้อย่างสะดวกและมีปริมาณมากแต่จะทราบได้อย่างไรว่าเนื้อหาบนเว็บไซต์นั้นมีความน่าเชื่อถือหรือไม่ งานวิจัยนี้จึงมีวัตถุประสงค์ในการจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ด้านมะเร็ง เพื่อแยกประเภทของเนื้อหาเว็บไซต์ที่มีความน่าเชื่อถือและไม่น่าเชื่อถือ ซึ่งงานวิจัยนี้นำเสนอ CancerDic+ เพื่อใช้ในการสกัดคำ โดยมีการเพิ่มข้อมูลคำศัพท์เฉพาะด้านเกี่ยวกับมะเร็งและใช้เหมืองข้อมูล (Text Mining) ทำการจำแนกข้อมูล โดยมีการเปรียบเทียบค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความครบถ้วน (Recall) ของการจำแนกความน่าเชื่อถือของเนื้อหาที่ผ่านเครื่องมือสกัดคำจาก Lexto SWATH และ CancerDic+ ซึ่งผลการจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์พบว่าการทำเหมืองข้อมูลโดยใช้ CancerDic+ สกัดคำให้ผลการจำแนกได้ดีที่สุด (Accuracy = 0.844, Precision = 0.838, Recall =0.845) ซึ่งสามารถนำไปประยุกต์ใช้งานอื่นได้อย่างมีประสิทธิภาพ

คำสำคัญ – เหมืองข้อความ; การจำแนกข้อมูล; การสกัดคำ; ความน่าเชื่อถือของเว็บไซต์

1. บทนำ

ปัจจุบันการเติบโตของข้อมูลบนอินเทอร์เน็ตเพิ่มขึ้นเป็นอย่างมาก ทำให้การค้นหาข้อมูลที่เกี่ยวข้องกับการแพทย์และการดูแลรักษาสุขภาพสามารถทำได้สะดวกและรวดเร็ว แต่ข้อมูลออนไลน์ที่อยู่บนอินเทอร์เน็ตมีอยู่เป็นจำนวนมาก ซึ่งมีทั้งข้อมูลที่มีความน่าเชื่อถือและไม่น่าเชื่อถือ จึงไม่สามารถนำข้อมูลเหล่านั้นมาใช้งานได้อย่างเหมาะสมและมีประสิทธิภาพได้ ซึ่งในการดูแลรักษาสุขภาพต้องการข้อมูลที่มีความน่าเชื่อถือเพื่อผู้ที่มาสืบค้นข้อมูลจะได้ปฏิบัติอย่างถูกวิธีป้องกันการเกิดผลกระทบกับร่างกายหรือโรคที่กำลังเป็นอยู่ ซึ่งในประเทศไทยนั้นคนไทยจะเสียชีวิตด้วยโรคร้ายแรงเพียงไม่กี่โรค ได้แก่ โรคมะเร็ง โรคหัวใจ โรคเบาหวาน โรคความดันโลหิตสูง โรคหลอดเลือดในสมอง โรคปอดอักเสบและโรคอื่นๆเกี่ยวกับปอด ในงานวิจัยนี้จึงมุ่งเน้นไปใช้เนื้อหาเกี่ยวกับโรคมะเร็งเนื่องจากเป็นโรคสำคัญหมายถึงโรคที่สามารถป้องกันได้แต่มีผู้เสียชีวิตเพิ่มขึ้นทุกปี ปัจจุบันมีอัตราการป่วยเพิ่มขึ้นอย่างต่อเนื่องและที่เป็นโรคที่มีอัตราการเสียชีวิตเป็นอันดับ 1 ตั้งแต่ปี พ.ศ. 2551-2556 จะเห็นได้จากรายงานสถิติข้อมูลของสำนักนโยบายและยุทธศาสตร์ กระทรวงสาธารณสุข แสดงดังรูปที่ 1 ดังนั้นหากมีข้อมูลในการดูแลรักษาเกี่ยวกับโรคมะเร็งในลักษณะออนไลน์ที่มีความน่าเชื่อถือก็จะสามารถให้ความรู้แก่ผู้ป่วยหรือผู้มาสืบค้นห้อมูล เพื่อช่วยลดอัตราการป่วยและการเสียชีวิตจากโรคมะเร็งได้



รูปที่ 1. แสดงจำนวนและอัตราผู้เสียชีวิตจากโรคสำคัญ ปี พ.ศ. 2551 – 2556

ดังนั้นการที่จะนำข้อมูลที่ได้จากเว็บไซต์มาใช้งาน จึงควรมีการจำแนกความน่าเชื่อถือของเว็บไซต์ ซึ่งในต่างประเทศมีการรับรองเนื้อหาภายในเว็บไซต์เกี่ยวกับทางการแพทย์และสุขภาพโดยมูลนิธิฮอน (Health On the Net Foundation: HON) ซึ่งก่อตั้งในสหภาพยุโรป เมื่อปี ค.ศ.1996 โดยมีจุดกำเนิดมาจากข่าว “Shark cartilage to cure cancer!” บนเว็บไซต์ซึ่งแปลได้ว่า “ครีบปลาฉลามสามารถรักษามะเร็ง!” ซึ่งเมื่อข่าวนี้ได้แพร่ออกไปทำให้เกิดผลกระทบในวงกว้างทั้งมีการเพิ่มจำนวนการล่าปลาฉลามเพื่อเอาครีบมาขายให้กับผู้ที่เชื่อในข่าวนี้ และมีผู้ป่วยหลงเชื่อจนเลิกการรักษากับแพทย์แผนปัจจุบันทำให้เกิดผลเสียต่อตัวผู้ป่วยจนทำให้แพทย์ต้องออกมาเตือนผู้รับข่าวนี้ จึงทำให้ต้องมีองค์กรที่มากำกับดูแลเนื้อหาภายในเว็บไซต์ด้านสุขภาพ โดยดำเนินการส่งเสริมและแนะนำการนำข้อมูลออนไลน์ด้านสุขภาพที่มีประโยชน์เชื่อถือได้มาใช้งานได้เหมาะสมและมีประสิทธิภาพ โดยมูลนิธิ HON จะมีเกณฑ์การพิจารณาความน่าเชื่อถือของเว็บไซต์ 8 ข้อ ดังนี้ คุณสมบัติของผู้เขียน (Authoritative) ความสมบูรณ์ของบทความ (Complementarity) ความเป็นส่วนตัวของผู้ใช้งาน (Privacy) การแสดงแหล่งที่มา (Attribution) มีรองรับเรื่องการร้องเรียน (Justifiability) ความโปร่งใสของข้อมูล (Transparency) ระบุแหล่งเงินทุน (Financial Disclosure) และแยกส่วนเนื้อหาและโฆษณาอย่างชัดเจน (Advertising) เว็บไซต์ที่ได้รับการรับรองความน่าเชื่อถือจะได้รับสัญลักษณ์ภาพ HONcode เพื่อนำไปแสดงไว้ที่เว็บไซต์นั้น ๆ ซึ่งทำให้ผู้สืบค้นข้อมูลจากเว็บไซต์ที่ได้รับการรับรองความน่าเชื่อถือมีความมั่นใจในการนำข้อมูลเหล่านั้นมาใช้งานได้อย่างเหมาะสมและเกิดประโยชน์ [1] สำหรับในประเทศไทยนั้นยังไม่มีการรับรองเนื้อหาภายในเว็บไซต์ที่เกี่ยวกับสุขภาพ แต่มีการรับรองเนื้อหาเว็บไซต์ทางด้านพาณิชย์อิเล็กทรอนิกส์ ซึ่งจัดทำโดยกรมพัฒนาธุรกิจการค้า กระทรวงพาณิชย์ โดยเว็บไซต์ที่ทำพาณิชย์อิเล็กทรอนิกส์ที่ต้องการได้รับเครื่องหมายรับรองจะต้องทำการลงทะเบียนผ่านการรับรองความน่าเชื่อถือ กับกรมพัฒนาธุรกิจการค้าแล้วต้องผ่านเกณฑ์ที่กำหนดไว้ จึงจะได้รับสัญลักษณ์ภาพ DBD Verify แสดงดังรูปที่ 2

รูปที่ 2. แสดงตัวอย่างสัญลักษณ์ภาพ DBD Verify

ไปติดไว้บนหน้าเว็บไซต์ของตนเอง ทำให้ผู้ซื้อสินค้าออนไลน์เกิดความเชื่อมั่นในตัวเว็บไซต์มากยิ่งขึ้น แต่ในการรับรองเนื้อหาเว็บไซต์เกี่ยวกับสุขภาพด้านมะเร็งในประเทศไทยนั้นยังไม่มีการจำแนกหรือรับรองความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์

งานวิจัยนี้จึงมีวัตถุประสงค์ในการจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ด้านมะเร็ง เพื่อแยกประเภทของเนื้อหาเว็บไซต์ที่มีความน่าเชื่อถือและไม่น่าเชื่อถือ โดยนำเสนอ CancerDic+ ซึ่งเป็นพจนานุกรมคำศัพท์เฉพาะด้านมะเร็งเพื่อตัดคำภาษาไทย โดยนำข้อมูลมาจากเว็บไซต์ ซึ่งประกอบไปด้วยข้อมูลที่มีความน่าเชื่อถือ เช่น ข้อมูลที่ได้จากเว็บไซต์หรือสถาบันทางการแพทย์ หรือบล็อกของแพทย์ ส่วนข้อมูลที่ไม่น่าเชื่อถือ เช่น ข้อมูลการขายประกันสุขภาพ ข้อมูลการขายอาหารเสริม ข้อมูลสมุนไพรหรือยาที่โฆษณาสรรพคุณเกินจริง จากนั้นจึงนำมาเข้ากระบวนการสกัดข้อความ (Text Extraction) เพื่อตัดคำ ในส่วนการให้ค่าน้ำหนักคำเพื่อที่จะนำไปเป็นตัวแทนเอกสารใช้วิธี การหาค่าน้ำหนัก (Term Weighting: TF) และการหาค่าความถี่ผกผัน (Inverse Document Frequency: IDF) แล้วจึงนำมาจำแนกประเภทความน่าเชื่อถือของเว็บไซต์ด้านมะเร็งและทำการประเมิน เพื่อเปรียบเทียบประสิทธิภาพการตัดคำและการจำแนกเนื้อหา

ในงานวิจัยได้แบ่งเนื้อหาออกเป็นส่วน ดังนี้ ส่วนที่ 2 แนวคิดและวิธีการดำเนินงาน ส่วนที่ 3 ผลการทดลอง และส่วนที่ 4 บทสรุป

2. แนวคิดและวิธีดำเนินงาน

จากการศึกษาการจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ด้านมะเร็ง มีทฤษฎีที่เกี่ยวข้อง และมีกรอบวิธีดำเนินการดังนี้

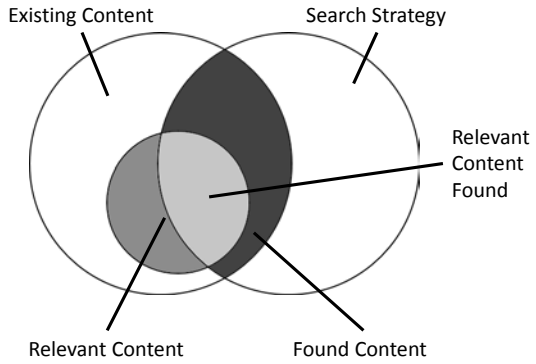
2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ความน่าเชื่อถือของเว็บไซต์ (Website Credibility)

นิยามของความน่าเชื่อถือ (Credibility) [1] หมายถึง ความเชื่อถือได้ (Believability) ไม่ว่าจะเป็นบุคคลหรือวัตถุซึ่งมีลักษณะที่เชื่อถือได้ 2 ประการ คือ ความรู้สึกว่ามีคุณภาพผู้คนรับรู้ว่ามีคุณภาพ (Perceived) ซึ่งอาจไม่มีอยู่ในตัวบุคคลหรือวัตถุสารสนเทศจริง และความน่าเชื่อถือที่ได้จากการรับรู้ (Perception of Credibility) การประเมินความน่าเชื่อถือเป็นผลมาจากสมองที่จะประเมินปัจจัยที่สำคัญ ได้แก่ ความไว้เนื้อเชื่อใจได้ (Trustworthiness) สำหรับการประเมินสารสนเทศต่าง ๆ ผ่านเว็บ และความเป็นผู้ชำนาญการ (Expertise) ซึ่งจะต้องมีความรู้ประสบการณ์และสมรรถนะ มีชื่อนักเขียนบทความและการอ้างอิงชัดเจน เป็นต้น ความไว้เนื้อเชื่อใจจะบอกถึงความดีและมีจรรยาบรรณของเว็บไซต์ ดังนั้นหากต้องการให้เว็บไซต์ที่มีความน่าเชื่อถือจะต้องทำให้ผู้มาเยี่ยมชมรับรู้ว่ามีความไว้เนื้อเชื่อใจได้ และความเป็นผู้ชำนาญการในระดับสูง นอกจากนี้ปัจจัย ที่ทำให้เว็บไซต์น่าเชื่อถือได้เพิ่มขึ้น เช่น การเพิ่มคุณค่าให้กับเว็บไซต์ โดยการปรับปรุงเนื้อหาให้ทันสมัย บทความต้องมีอ้างอิงหรือผู้แต่งเสมอไม่มีโฆษณามากเกินไป นามสกุลของเว็บไซต์ (Domain Name) ต้องเป็นขององค์กรที่จดทะเบียนอย่างถูกต้อง ทุกองค์ประกอบบนเว็บไซต์ทำงานได้ถูกต้องและชื่อเสียงในด้านดีขององค์กรก็จะส่งผลต่อเว็บไซต์ด้วย

2.1.2 ทฤษฎีด้านเซ็ตสำหรับการสืบค้นสารสนเทศ

ในเรื่องของเซตนั้นมักจะถูกนำใช้กับการอธิบายวัตถุนามธรรมให้เข้าใจได้ง่ายในรูปแบบของวัตถุทางคณิตศาสตร์ เพื่อทำให้เรื่องที่อธิบายได้ยากสามารถเข้าใจได้โดยง่าย ซึ่งเครื่องมืออย่างแผนภาพเวนน์ก็ถูกนำมาใช้ในการอธิบายเรื่องเซตอยู่บ่อยครั้ง เนื่องจากทำให้มองเห็นเป็นภาพและอธิบายความหมายได้ดี ซึ่งในการสืบค้นสารสนเทศก็สามารถนำมาอธิบายได้ด้วยเซตเช่นกัน โดยมองเนื้อหาที่ต้องการค้นหาบนอินเทอร์เน็ตเป็นเซตใหญ่ทั้งหมด (Existing Content) แล้วจะมีเพียงเนื้อหาบนอินเทอเน็ตบางอย่างเท่านั้นที่ตรงประเด็นที่ผู้ค้นหาต้องการ (Relevant Content) ซึ่งมีเซตของกลยุทธ์ในการสืบค้น (Search Strategy) ที่จะเป็นตัวกำหนดว่าผู้ค้นข้อมูลจะใช้วิธีการในการสืบค้นอย่างไรให้ได้ข้อมูลที่ต้องการได้ในปริมาณมาก ซึ่งในสิ่งที่ผู้ค้นข้อมูลได้รับออกมาอาจจะมีทั้งเนื้อหาที่มีอยู่บนอินเทอร์เน็ตหรือไม่ได้อยู่บนอินเทอร์เน็ตก็ได้ ทำให้เกิดเซตทั้งสองส่วนที่ซ้อนทับกัน (Intersect) คือส่วนของเนื้อหาที่ถูกค้นพบจากการสืบค้น (Found Content) ซึ่งเซตที่จะใช้งานได้จริงหรือตรงประเด็นที่ผู้ค้นข้อมูลจะใช้ได้จริงจะเล็กลงไปอีก แสดงดังรูปที่ 3



รูปที่ 3. แสดงเซตการสืบค้นสารสนเทศ

จากรูปจะสรุปได้ว่าข้อมูลที่ผู้ใช้งานต้องการค้นหานั้นจะมีอยู่จำนวนหนึ่งซึ่งจะได้ข้อมูลที่มากหรือน้อย ขึ้นอยู่กับกลยุทธ์ในการค้นหาข้อมูลของผู้ใช้งาน และในส่วนของข้อมูลที่ค้นหามาแล้วนั้นถ้าเป็นข้อมูลที่เกี่ยวกับสุขภาพ ก็จะต้องใช้ข้อมูลที่มีความน่าเชื่อถือ ซึ่งข้อมูลที่ค้นหามาได้นั้นจะมีทั้งข้อมูลที่น่าเชื่อถือไม่น่าเชื่อถือปะปนกันอยู่ ทำให้สิ่งที่ค้นหามาได้นั้นสามารถนำมาใช้งานได้จริงๆมีจำนวนน้อยลงไปอีก และถ้าไม่คำนึงถึงความน่าเชื่อถือของเนื้อหาที่ค้นหามาได้ ก็อาจจะส่งผลเสียต่อผู้ที่นำไปปฏิบัติตามข้อมูลที่ไม่ถูกต้อง

2.1.3 การสกัดข้อความ (Text Extraction)

การสกัดข้อความนั้นเป็นกระบวนการของการทำเหมืองข้อความ (Text Mining) เพื่อใช้การวิเคราะห์คำออกจากเอกสาร ข่าวสาร ข้อความ และสารสนเทศต่าง ๆ ที่เป็นตัวอักษรโดยสามารถนำไปทำการแบ่งกลุ่ม (Clustering) การจำแนกข้อมูล (Classification) และการหาความสัมพันธ์ (Association) ซึ่งในการแบ่งกลุ่มเอกสาร (Document Clustering) [2] เป็นการวัดความคลายคลึงกันของข้อความในตัวเอกสาร โดยข้อมูลตัวอักษรจะถูกแปลงเป็นตัวเลขเพื่อทำขั้นตอน การแบ่งกลุ่มโดยใช้เทคนิคต่างๆ เช่น DBSCAN, K-mean, SOM และ Hierarchical ซึ่งก่อนการทำเหมืองข้อความจะต้องผ่านขั้นตอนการเตรียมข้อมูล (Preprocess) ก่อนซึ่งมีขั้นตอน [3][4][5] ดังนี้

2.1.3.1 การตัดคำ (Word Segmentation)

เป็นการแยกแต่ละคำจากเอกสารออกจากกัน โดยยังคงมีความหมายที่ถูกต้องสมบูรณ์อยู่ โดยการตัดคำนั้นใช้ฐานข้อมูลพจนานุกรมคำศัพท์ ในการแบ่งคำออกมา [6]

2.1.3.2 เทคนิคการตัดคำ (Word Segmentation Techniques)

ในการตัดคำมีเทคนิคที่สามารถนำมาใช้โดยขึ้นอยู่กับลักษณะของคำที่จะนำมาตัดซึ่งมีเทคนิคที่นิยมใช้งานกันแบ่งออกเป็น 5 เทคนิค ดังนี้ เทคนิคการเทียบคำที่ยาวที่สุด (Longest Word Pattern Matching) เทคนิคการเทียบคำที่สั้นที่สุด (Shortest Word Pattern Matching) เทคนิคการตัดคำที่ใช้ความถี่ของคำหรือสถิติ (Probabilistic Word Segmentation) เทคนิคการย้อนรอยกลับ (Back Tracking) และเทคนิคการตัดคำแบบใช้คุณลักษณะ (Feature-based Approach) [7]

งานวิจัยเกี่ยวกับการตัดคำมีอยู่มากมาย ทางผู้วิจัยได้ทำการศึกษางานวิจัยของวิโรจน์ [8] เสนอการตัดคำด้วยวิธีไตรแกรมกับคลังข้อมูลเพื่อตัดคำระดับพยางค์และทดสอบกับพจนานุกรมแต่วิธี Maximum Collocation ไม่สามารถแบ่งคำได้ชัดเจนขึ้นอยู่กับคำที่มีอยู่ในพจนานุกรม และควรใช้วิธีการทางสถิติอื่นมาปรับปรุงประสิทธิภาพ ปโยธร [9] ใช้เทคนิควิธีการตรวจสอบย้อนกลับ (Back Tracking) และการเลือกคำที่ยาวที่สุด (Longest Matching) และควรให้โปรแกรมเพิ่มคำเฉพาะที่ไม่พบในพจนานุกรมในตอนแรกโดยอัตโนมัติเพื่อให้การตัดคำในเอกสารแบบเดียวกันเป็นไปได้ดีขึ้น สิทธิโชค [10] สร้างโมเดลวิธีการตัดคำภาษาไทยด้วยเทคนิคการเรียนรู้ของเครื่องพบปัญหาการคำนวณความถูกต้องและแม่นยำของคอมพิวเตอร์รวมทั้งการจัดขอบขวาในโปรแกรมประมวลผลคำเพื่อค้นหาคำส่วนของการแบ่งคำจะเสียเวลาในขั้นตอนการแปลงไฟล์หลายไฟล์เพื่อรวมเป็นไฟล์เดียวก่อนที่จะนำมาเป็นข้อมูลสำหรับสอน (Train Data) หรือข้อมูลสำหรับทดสอบ (Test Data) และมีข้อจำกัดคือสามารถตัดคำภาษาไทยที่มีข้อความเฉพาะอักขระเท่านั้น สำหรับงานวิจัยที่เกี่ยวข้องกับการตัดคำเพื่อแก้ปัญหาคำกำกวมโดย ชนินทร์ [11] ตัดคำด้วยวิธีพจนานุกรมคำกำกวมเพื่อแก้ปัญหาคำกำกวมและคำที่ไม่ปรากฏในพจนานุกรมด้วยโมเดล PTTSF (Parsing Thai Text with Syntax and Feature of Word) พบปัญหาไม่สามารถตัดคำที่เป็นคำกริยาระหว่างคำที่ไม่ปรากฏในพจนานุกรมได้ กานดา [2] ตัดคำในเอกสารภาษาไทยโดยการใช้กฎ (Rule-based) และพจนานุกรมแบบใหม่ร่วมกันมีการตัดคำในระดับพยางค์ ส่วนคำกำกวมสามารถแก้ปัญหาโดยการใช้วิธีตัดคำที่ยาวที่สุดร่วมกับวิธีการย้อนกลับแต่ยังไม่สามารถตัดคำได้ถูกต้องทุกครั้ง และควรมีการเพิ่มเติมปรับปรุงสำหรับการตัดคำประเภทนี้ด้วยการใช้ค่าสถิติของคำที่พบใน

เอกสารทั่วไปร่วมกับการใช้ความถี่ของคำที่พบในเอกสารที่นำมาตัดคำ ชูชาติ [12] นำเสนอกรอบการทำงานการเก็บคำที่ไม่รู้จักจากเว็บ โดยใช้การวิเคราะห์คำที่ไม่รู้จักร่วมกับพจนานุกรมเพื่อสกัดคำที่ไม่รู้จักโดยอัตโนมัติ ทำให้ผู้ใช้สามารถเพิ่มคำไทยที่ไม่รู้จักลงในพจนานุกรมคำไทยที่ไม่รู้จัก ผลการทดลองพบว่าสามารถวิเคราะห์คำที่ไม่รู้จักได้สูงสุดถึงร้อยละ 96

2.1.4 การจำแนกประเภทข้อมูล (Data Classification)

การทำเหมืองข้อความเป็นการสกัดเอาสิ่งที่มีประโยชน์ออกมาจากข้อความที่มีจำนวนมาก ซึ่งมีหลากหลายวิธี หนึ่งในนั้นคือการจำแนกข้อมูล (Classification) ซึ่งมีเทคนิคต่าง ๆ ที่นิยมใช้งานดังนี้

2.1.4.1 ต้นไม้ตัดสินใจ (Decision Tree)

เทคนิคต้นไม้ตัดสินใจ เป็นวิธีหนึ่งที่ใช้ในการจำแนกข้อมูล โดยมีลักษณะการทำงานเหมือนโครงสร้างของต้นไม้ รูปแบบของต้นไม้ตัดสินใจประกอบด้วยโหนดราก (Root Node) ซึ่งเป็นโหนดแรก และแตกสายย่อยเป็นโหนดลูก (Child Node) [13][14] โดยจะสามารถแปลงไปเป็นกฎที่ใช้ในการจำแนกข้อมูลหรือที่เรียกว่าฐานกฎ (Rule-based) เพื่อใช้ในการพยากรณ์ข้อมูล โดยกฎนั้นจะเป็นไปตามรูปแบบของต้นไม้ตัดสินใจ

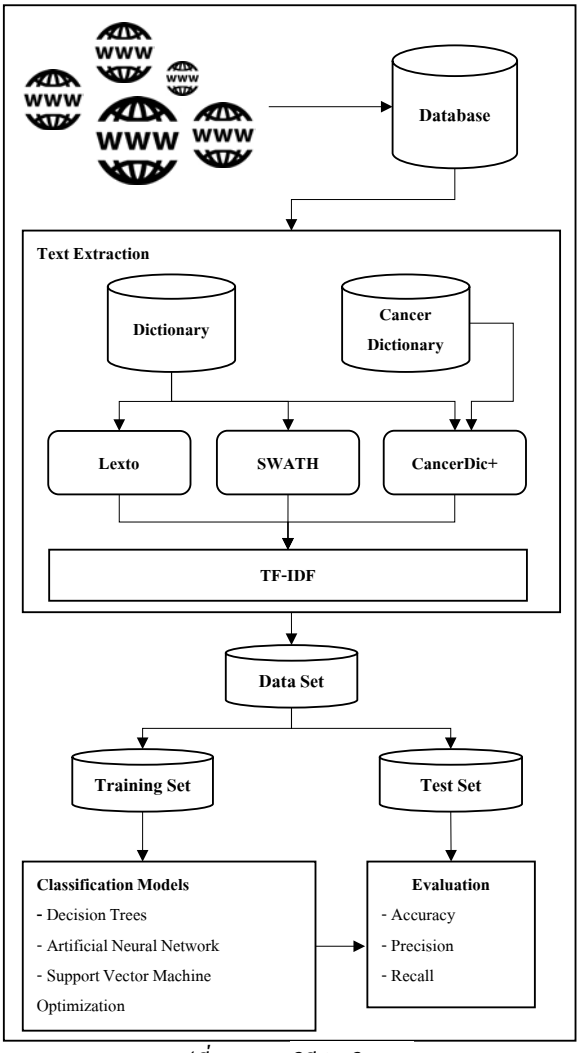
2.1.4.2 โครงข่ายประสาทเทียม (Artificial Neural Network: ANN)

เป็นการจำแนกข้อมูลอีกวิธีหนึ่งที่นิยมใช้งาน มีการทำงานเลียนแบบหลักการทำงานของสมองมนุษย์ โดยมีหน่วยที่ใช้ในการประมวลผลเรียกว่า นิวรอน ซึ่งนิวรอนแต่ละนิวรอนสามารถรับค่าได้หลายอินพุตแต่มีเอาต์พุตได้เพียงเอาต์พุตเดียวเท่านั้น โดยทุกอินพุตมีค่าถ่วงน้ำหนัก (Weight) และในแต่ละนิวรอนนั้นจะมีค่าความเอนเอียงหรือไบแอส (Bias) อยู่หรือไม่ก็ได้ โดยเมื่อปรับค่าถ่วงน้ำหนักและค่าเอนเอียงที่เหมาะสมแล้วจะถูกส่งไปยังฟังก์ชันถ่ายโอน (Transfer Function) [15][16] เพื่อคำนวณค่าผลลัพธ์

2.1.4.3 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM)

เป็นเทคนิคที่มีการใช้สมการทางคณิตศาสตร์ในการจำแนกข้อมูล โดยพยายามหาจุดที่เส้นแบ่งกลุ่มของข้อมูลมีระยะความห่างระหว่างเส้นขอบเขต (Border Line) มากที่สุด ซึ่งวิธีนี้ทำให้มีข้อดี คือ สามารถรองรับจำนวนตัวแปรที่หลากหลายได้เป็นจำนวนมาก และค่อนข้างมีความถูกต้องสูง ซึ่งในการทำงานของตัวซัพพอร์ตเวกเตอร์แมชชีน มีฟังชั้นให้เลือกใช้อย่างหลากหลาย เช่น Linear Function, Polynomial Function และRadial Basis Function [17] แต่ต้องมีการเลือกใช้ฟังชั่นให้ตัวซัพพอร์ตเวกเตอร์แมชชีน จำแนกข้อมูลได้อย่างเหมาะสมด้วยเช่นกัน

2.2 วิธีดำเนินการ



รูปที่ 4. แสดงวิธีดำเนินงาน

จากรูปที่ 4 แสดงวิธีดำเนินงานวิจัยการจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ด้านมะเร็งโดยใช้ CancerDic+ ซึ่งการสกัดคำผ่านเครื่องมือ CancerDic+ เป็นการตัดคำโดยใช้เทคนิคเดียวกับ Lexto โดยใช้เทคนิคการเทียบคำที่ยาวที่สุด (Longest Word Pattern Matching) ร่วมกับวิธีการใช้พจนานุกรม (Dictionary-based) โดยเพิ่มพจนานุกรมคำศัพท์เฉพาะด้านมะเร็ง ในส่วนของการตัดคำเพื่อให้ผลของการตัดคำที่เป็นเนื้อหาที่มีคำศัพท์ด้านมะเร็งอยู่มากมีความถูกต้องเพิ่มมากขึ้น จึงได้เตรียมข้อมูลสำหรับการจำแนกโดยมีขั้นตอนดังนี้

2.2.1 การเก็บรวบรวมข้อมูล (Data Collection)

การเก็บข้อมูลโดยใช้ข้อมูลจากเว็บไซต์ที่สืบค้นด้วยคำสำคัญที่เกี่ยวข้องกับมะเร็ง เช่น มะเร็ง รักษามะเร็ง อาการของมะเร็ง อาหารสำหรับผู้ป่วยมะเร็ง อาหารเสริมสำหรับมะเร็ง สมุนไพรรักษามะเร็ง เป็นต้น โดยมีจำนวน 484 แถว ซึ่งสามารถแบ่งขอบเขตข้อมูลออกเป็น 6 ส่วน แสดงดังตารางที่ 1 ในส่วนของหลักเกณฑ์ที่ใช้ในการกำกับประเภทของเนื้อหาในเว็บไซต์ว่าเป็นเนื้อหาที่มีความน่าเชื่อถือหรือไม่น่าเชื่อถือ นั้นใช้จากแหล่งที่มาของข้อมูลหรือที่อยู่ของเว็บไซต์ ซึ่งข้อมูลเนื้อหาที่น่าเชื่อถือมาจากเว็บไซต์ประเภท โรงพยาบาลรัฐบาลและเอกชน หน่วยงานด้านสาธารณสุขของภาครัฐ หรือบล็อกของแพทย์ ส่วนข้อมูลเนื้อหาที่ไม่น่าเชื่อถือมาจากเว็บไซต์ประเภท ขายประกัน ขายอาหารเสริมที่มีสรรพคุณเกินจริง และข้อมูลจากเว็บไซต์ทั่วไปที่ไม่มีการอ้างอิงที่มาของเนื้อหา

ตารางที่ 1. แสดงขอบเขตและรูปแบบของข้อมูลที่จัดเก็บในฐานข้อมูล

#	Name	Detail	Type
1	Text	เนื้อหาในเว็บไซต์	Text
2	Lexto	ข้อความที่สกัดผ่าน Lexto	Text
3	SWATH	ข้อความที่สกัดผ่าน SWATH	Text
4	Dic	ข้อความที่สกัดผ่าน CancerDic+	Text
5	Link	ลิงค์ที่มาของเนื้อหาในเว็บไซต์	Text
6	Type	ประเภทของเนื้อหา (1 = น่าเชื่อถือ, 2= ไม่น่าเชื่อถือ)	Char(1)

2.2.2 การสกัดคำ (Text Extraction)

ในการสกัดคำภาษาไทยมีเครื่องมือสำหรับตัดคำอยู่หลายเครื่องมือ แต่ในงานวิจัยนี้เลือกใช้เครื่องมือสำหรับการตัดคำ 3 เครื่องมือ ได้แก่ 1) เล็กซ์โต (Thai Lexeme Tokenizer: LexTo) ที่ถูกพัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ 2) SWATH (Smart Word Analysis for THai) ที่ถูกพัฒนาโดยมหาวิทยาลัยคาร์เนกีเมลลอน และ 3) CancerDic+ (Cancer Dictionary Plus) ที่ผู้วิจัยนำเสนอ โดยเมื่อผ่านขั้นตอนการตัดคำ (Word Segmentation) แล้วจะนำไปบันทึกลงในฐานข้อมูลที่ผ่านการสกัดคำจากเครื่องมือทั้ง 3 ดังกล่าวข้างต้น เพื่อเป็นการเตรียมข้อมูลสำหรับการสอน (Train Data) และข้อมูลสำหรับการทดสอบ (Text Data) ในขั้นตอนถัดไป

2.2.3 การสร้างดัชนีเอกสาร (Document Indexing)

การสร้างดัชนีเอกสารเป็นขั้นตอนการแปลงเอกสารซึ่งเป็นภาษาธรรมชาติให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถประมวลผลได้ ซึ่งจะเป็นการสร้างตัวแทนเนื้อหาเอกสารให้อยู่ในรูปแบบเวกเตอร์ของน้ำหนักคำ (Term Weighting) เพื่อนำมาหาค่าสร้างดัชนี โดยนิยมใช้รูปแบบของคำเดี่ยว [13] เริ่มจากการสร้างเวกเตอร์ตัวแทนเอกสาร จากนั้นจะสร้างเมตริกซ์ของเอกสารขึ้นจากเวกเตอร์เอกสารทั้งหมด ในกลุ่ม จนกระทั่งกลายเป็นเมตริกซ์

ในส่วนของการคำนวณค่าน้ำหนักให้แก่ดัชนีในแต่ละเอกสารใช้วิธีการ TF-IDF Weighting (Term Frequency-Inverse Document Frequency) แสดงดังสมการที่ (1)(2)(3)(4) [16]

$$TF - IDF = TF \times IDF \quad (1)$$

$$TF_t = \frac{n_t}{N} \quad (2)$$

$$IDF_t = 1 + \log\left(\frac{D}{d_t}\right) \quad (3)$$

$$TF - IDF_t = \frac{n_t}{N} \times \left[1 + \log\left(\frac{D}{d_t}\right)\right] \quad (4)$$

โดยที่

n_t = จำนวนคำ t ที่ปรากฏในเอกสาร

N = จำนวนคำทั้งหมดที่ปรากฏในเอกสาร

D = จำนวนเอกสารทั้งหมด

d_t = จำนวนเอกสารที่มีคำ t ปรากฏ

2.2.4 การจำแนกข้อมูลและการวัดประสิทธิภาพ

ในงานวิจัยนี้ผู้วิจัยได้ทำการเปรียบเทียบค่าประสิทธิภาพซึ่งพิจารณาจากค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และความครบถ้วน (Recall) จากโมเดล 3 รูปแบบ แสดงดังตารางที่ 2 โดยทุกโมเดลใช้วิธีการแบ่งชุดข้อมูลเพื่อใช้ในการสอนและทดสอบด้วยวิธี 10-fold Cross Validation เพื่อวัดความเที่ยงตรงของโมเดล

ตารางที่ 2. แสดงรายละเอียดโมเดลที่ใช้ในการทดลอง

โมเดล	ชื่อโมเดล	ชื่อย่อ
1	Decision Trees	DT
2	Artificial Neural Network (Back Propagation)	ANN
3	Support Vector Machine Optimization (Linear Function)	SMO

การหาค่าประสิทธิภาพนั้นใช้การพิจารณาค่าประสิทธิภาพจากระดับค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความครบถ้วน (Recall) แสดงดังสมการที่ (5)(6)(7) [8]

$$Precision(P) = \frac{TP}{TP+FP} \quad (5)$$

$$Recall(R) = \frac{TP}{TP+FN} \quad (6)$$

$$Accuracy(A) = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

โดยเมื่อ

TP = จำนวนของ True Positive

FP = จำนวนของ False Positive

FT = จำนวนของ True Negative

FN = จำนวนของ False Negative

3. ผลการทดลอง

ผู้วิจัยได้ดำเนินการจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ด้านมะเร็ง โดยทำการสกัดคำ การคำนวณหาค่าดัชนีเอกสาร และเปรียบเทียบประสิทธิภาพการจำแนกประเภทความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ มีรายละเอียด ดังนี้

3.1 ผลการสกัดคำและสร้างดัชนีเอกสาร

ในการดำเนินการสกัดคำจากข้อมูลเนื้อหาด้านมะเร็งด้วยเครื่องมือ LexTo และ SWATH กับฐานข้อมูลพจนานุกรมเล็กซิตรอน แต่ในส่วนของ CancerDic+ นั้นได้ใช้ฐานข้อมูลพจนานุกรมเล็กซิตรอนผสมกับฐานข้อมูลคำศัพท์เฉพาะทางด้านมะเร็ง โดยแสดงตัวอย่างการตัดคำด้วยโปรแกรมทั้ง 3 โปรแกรมข้างต้น แสดงดังรูปที่ 5

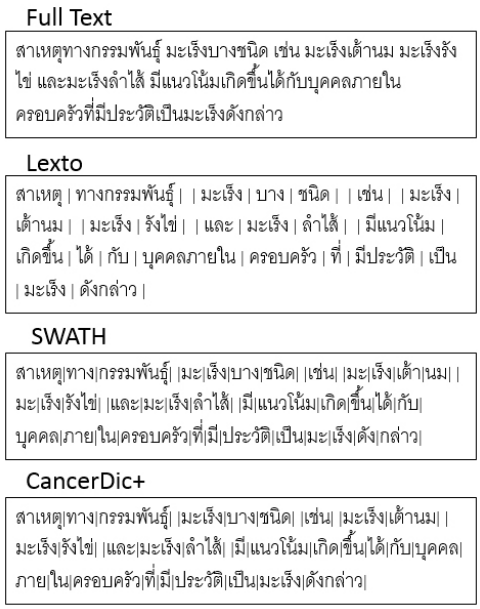
ซึ่งจากการใช้เครื่องมือทั้ง 3 ในการสกัดคำจะพบว่าโปรแกรม Lexto และ CancerDic+ มีการตัดคำที่เป็นคำสำคัญในด้านมะเร็งใกล้เคียงกัน แต่ในส่วนของโปรแกรม SWATH มีการตัดคำสำคัญในด้านมะเร็งที่แตกต่างออกไป เช่น

“มะเร็ง” ผลที่ได้ Lexto และ CancerDic+ ตัดคำได้ออกมาเป็น |มะเร็ง| แต่ SWATH ตัดคำออกมาเป็น |มะ|เร็ง|

“เต้านม” ผลที่ได้ Lexto และ CancerDic+ ตัดคำได้ออกมาเป็น |เต้านม| แต่ SWATH ตัดคำออกมาเป็น |เต้า|นม|

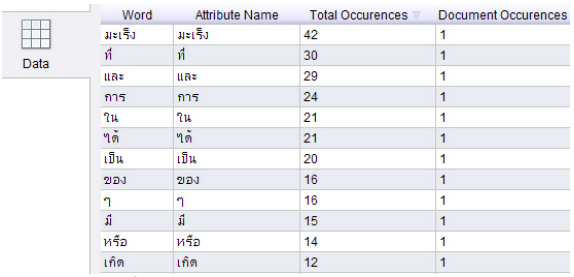
“ลำใส้” ผลที่ได้ Lexto SWATH และ CancerDic+ ตัดคำได้ออกมาเป็น |ลำใส้| เหมือนกัน

“ทางกรรมพันธุ์” ผลที่ได้ SWATH และ CancerDic+ ตัดคำได้ออกมาเป็น |ทาง|กรรมพันธุ์| แต่ Lexto ตัดคำออกมาเป็น |ทางกรรมพันธุ์|



รูปที่ 5. แสดงตัวอย่างผลการตัดคำ

ซึ่งจะเห็นว่ามีความแตกต่างกันในการตัดคำที่ผ่านโปรแกรมตัดคำ ซึ่งจะส่งผลกับประสิทธิภาพในการสร้างดัชนีเอกสารและจำแนกข้อมูล และในส่วนของการสร้างดัชนีเอกสารได้ใช้โปรแกรม RapidMiner Studio ช่วยในการนับคำและคำนวณ TD-IDF ซึ่งมีตัวอย่างการสร้างดัชนีเอกสาร ซึ่งจะถูกนำไปใช้ในการจำแนกข้อมูล แสดงดังรูปที่ 6


Data

Word	Attribute Name	Total Occurences	Document Occurences
มะเร็ง	มะเร็ง	42	1
ที่	ที่	30	1
และ	และ	29	1
การ	การ	24	1
ใน	ใน	21	1
ได้	ได้	21	1
เป็น	เป็น	20	1
ของ	ของ	16	1
ๆ	ๆ	16	1
มี	มี	15	1
หรือ	หรือ	14	1
เกิด	เกิด	12	1

รูปที่ 6. แสดงตัวอย่างการสร้างดัชนีเอกสาร

3.2 ผลการหาค่าประสิทธิภาพโดยใช้ Lexto, SWATH และ CancerDic+ สกัดคำ

จากการทดลองจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์โดยใช้โมเดลการจำแนก 3 แบบได้แก่ ต้นไม้ตัดสินใจ (DT) โครงข่ายประสาทเทียม (ANN) และซัพพอร์ตเวกเตอร์แมชชีน (SMO) โดยผ่านการสกัดคำโดยใช้โปรแกรม Lexto, SWATH และ CancerDic+ โดยการแบ่งข้อมูลเพื่อใช้ในการทดสอบความเที่ยงตรงของโมเดลด้วยวิธี 10-fold Cross Validation ทำให้ได้ค่าประสิทธิภาพ ซึ่งประกอบไปด้วยค่าความถูกต้อง ค่าความแม่นยำ และค่าความครบถ้วน แสดงดังตารางที่ 3

ตารางที่ 3. แสดงการเปรียบเทียบค่าประสิทธิภาพโดยใช้ Lexto, SWATH และ CancerDic+ สกัดคำ

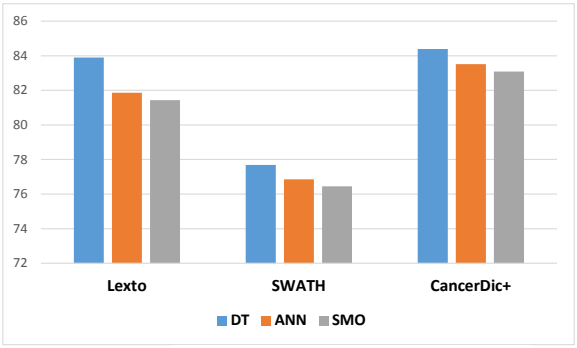
Tools	Model	10-fold cross validation		
		Accuracy	Precision	Recall
Lexto	DT	0.839	0.839	0.839
	ANN	0.819	0.814	0.819
	SMO	0.814	0.810	0.815
SWATH	DT	0.777	0.775	0.783
	ANN	0.769	0.765	0.773
	SMO	0.764	0.762	0.769
CancerDic+	DT	0.844	0.838	0.845
	ANN	0.835	0.829	0.836
	SMO	0.831	0.825	0.830

จากตารางพบว่าการตัดคำที่ผ่านเครื่องมือ Lexto แล้วมีค่าประสิทธิภาพมากที่สุด คือ DT (Accuracy = 0.839, Precision = 0.839, Recall =0.839) การตัดคำที่ผ่านเครื่องมือ SWATH แล้วมีค่าประสิทธิภาพมากที่สุด คือ DT (Accuracy = 0.777, Precision = 0.775, Recall =0.783) และการตัดคำที่ผ่านเครื่องมือ CancerDic+ แล้วมีค่าประสิทธิภาพมากที่สุด คือ DT (Accuracy = 0.844, Precision = 0.838, Recall =0.845) ซึ่งจะพบว่าโมเดลที่มีประสิทธิภาพมากที่สุดเป็นโมเดลการจำแนกข้อมูลแบบต้นไม้ตัดสินใจ (Decision Trees) ที่ผ่านการตัดคำจากเครื่องมือ CancerDic+

4. บทสรุป

จากการทดลองหาค่าประสิทธิภาพการจำแนกจากทั้ง 3 โมเดลจากการสกัดคำจากทั้ง 3 โปรแกรมทำให้ได้ค่าประสิทธิภาพของแต่ละโมเดลในแต่ละการทดลองจึงนำมาเปรียบเทียบกันทั้ง 3 รูปแบบการทดลองเพื่อให้เห็นภาพได้อย่างชัดเจน โดยใช้ค่าความถูกต้อง (Accuracy) มาทำกราฟเปรียบเทียบ แสดงดังรูปที่ 7

จากรูปพบว่าผลการทดลองโมเดลที่มีค่าประสิทธิภาพในด้านต่างๆ ที่มากที่สุด คือ โมเดลที่ผ่านการสกัดคำจาก CancerDic+ และใช้เทคนิคการจำแนกแบบ DT (Accuracy = 0.844, Precision = 0.838, Recall =0.845) ซึ่งดีกว่าโมเดลที่ผ่านการสกัดคำจาก Lexto และ SWATH เนื่องจากมีการใช้คำศัพท์เฉพาะทางเกี่ยวกับโรคมะเร็ง (Cancer Terminology) ทำให้การสกัดคำในขั้นตอนการเตรียมข้อมูลก่อนนำไปทำเหมืองข้อมูลมีความถูกต้องมากยิ่งขึ้น



รูปที่ 7. แสดงเปรียบเทียบค่าความถูกต้อง (Accuracy)

ในส่วนของปัญหาและข้อเสนอแนะ ที่พบจากการเก็บรวบรวมข้อมูลเนื้อหาจากเว็บไซต์ต่างๆ พบปัญหาที่ส่งผลต่อประสิทธิภาพการตัดคำ เช่น การสะกดคำผิด การใช้คำทับศัพท์ภาษาอังกฤษ ข้อมูลที่เป็นการให้ข้อมูลเชิงรูปภาพ (Info-graphic) อีกทั้งในการเก็บรวบรวมข้อมูลควรพัฒนาการเก็บข้อมูลโดยการใช้เว็บครอว์เลอร์ (Web Crawler) เพราะจะทำให้สามารถรวบรวมข้อมูลได้มาก รวดเร็วและหลากหลาย ส่วนกำหนดเกณฑ์อื่นๆ ที่ใช้ในการประเมินค่าความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ ที่ในงานวิจัยนี้ใช้เพียงแหล่งที่มาของเว็บไซต์ อาจจะใช้เนื้อหาประกอบกับชื่อผู้เขียน หรือประเมินจากผู้เชียวชาญก็ได้ ซึ่งจะนำไปพัฒนาต่อไปในอนาคต

เอกสารอ้างอิง

[1] “What makes a website credible?.” (ออนไลน์). แหล่งที่มา http://captology.stanford.edu/resources/what-makes-a-website-credible.html

[2] นิเวศน์ จิระวิชิตชัย. “แบบจำลองการจำแนกเอกสารสำหรับภาษาไทยอัตโนมัติ,” *The Journal of Industrial Technology* 2556, Vol. 2556. No. 1.

[3] X. Dai, Y. He, and Y. Sun, “A Two-layer Text Clustering Approach for Retrospective News Event Detection,” *International Conference on Artificial Intelligence and Computational Intelligence (AICI)*, vol. 1, pp. 364–368, 2010.

[4] U. Gunasinghe, S. Matharage, and D. Alahakoon, “A sequence based dynamic SOM model for text clustering,” *International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, 2012.

[5] K.Norvag and R.Oyri, “New Item Extraction for Text Mining in Web Newspapers,” *International Workshop on Challenges in Web Information Retrieval and Integration*, 2005.

[6] นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์ และพยุง มีสัจ. “การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ,” *NIDA Development Journal*, Vol. 51, No. 3, หน้า 187-205, 2554.

[7] สายัณห์ เทพแดง. “*การปรับปรุงประสิทธิภาพของการตัดคำไทย ด้วยเทคนิคการจดจำนิพจน์ระบุนาม*,” ปริญญานิพนธ์ วท.ม. สาขาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์, 2553.

[8] Wirote Aroonmanakun. “Collocation and Thai Word Segmentation,” *PROCEEDINGS OF SNLP-Oriental COCOSDA*, pp. 68-75, 2002.

[9] ปโยธร อุราธรรมกุล และ กานดา รุณนะพงศา. “การตัดคำภาษาไทยด้วยวิธีปรับปรุงกฎและพจนานุกรมแบบใหม่,” *JCSSE 2006*, vol. 2549, pp. 34-40, 2006.

[10] สิทธิโชค ทรัพย์ไพบูลย์กิจ และ สุพัฒนวรี ทิพย์เจริญ. “*การเพิ่มประสิทธิภาพการตัดคำภาษาไทยด้วยเทคนิคการเรียนรู้ด้วยเครื่อง*,”ปริญญานิพนธ์ วท.ม. ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่, 2549.

[11] ชนินทร์ มหัทธนชัย. “การพัฒนาเทคนิคการตัดคำแบบอาศัยไวยกรณ์และบริบทคำรอบข้าง,” *National Conference on Computer Information Technologies*, 2555.

[12] Choochart Haruechaiyasak. “A Collaborative Framework for Collecting Thai Unknown words from the web,” *Proceeding COLING-ACL'06,* pp. 345-352, 2006.

[13] มงคล ยูสเปรมานนท์. “การย่อความเอกสารภาษาไทย เชิงความหมายโดยใช้กราฟมโนภาพ,” วิทยานิพนธ์ ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และ เทคโนโลยี มหาวิทยาลัยธรรมศาสตร์, 159 หน้า, 2552.

[14] ธัญรัตน์ น้อมพลกรัง. “การพัฒนาระบบให้คำแนะนำ ตามรูปแบบ การปฏิสัมพันธ์ในกลุ่มการเรียนโดยใช้เทคนิค การทำเหมืองข้อมูลเพื่อส่งเสริมการเรียนรู้แบบร่วมมือ,” วิทยานิพนธ์ ภาควิชาคอมพิวเตอร์ศึกษา คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยพระจอมเกล้าพระนครเหนือ, 131 หน้า, 2557.

[15] ราชวิทย์ ทิพย์เสนา, ฉัตรเกล้า เจริญผล และแกมกาญจน์ สมประเสริฐศรี. “การจำแนกกลุ่มคำถามอัตโนมัติบนกระดานสนทนาโดยใช้เทคนิคเหมืองข้อความ,” J Sci Technol MSU, Vol 33, No. 5, หน้า -502, 2014.

[16] J. R. Quinlan. “Induction of Decision Trees,” *in Machine Learning*, pp. 81–106, 1986.

[17] Zlatko J. Kovcic. “Early Prediction of Student Success: Mining Students Enrolment Data,” *Proceeding of Informing Science & IT Education Conference (ImSITE)*, pp. 647-665, 2010.

[18] Edin Osmanbegovic, Mirza Suljic. “Data Mining Approach for Predicting Student Performance,” *Journal of Economics and Business*, Vol. 5, No. 1, pp. 3-12, 2012.

[19] พยุง มีสัจ. **ระบบฟัซซีและโครงข่ายประสาทเทียม**. พิมพ์ครั้งที่ 1. กรุงเทพฯ : ศูนย์ผลิตตำราเรียน มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2555.

[20] ภัทร์พงศ์ พงศ์ภัทรกานต์. “การวิเคราะห์ปัจจัยที่ส่งผลต่อการพ้นสภาพของนักศึกษาระดับปริญญาตรี โดยใช้คอมมิตตีแมชชีน,” *The 6th National Conference On Computing And Information Technology*, pp. 491-496, 2010.