การจำแนกกลุ่มคำถามอัตโนมัติบนกระดานสนทนา โดยใช้เทคนิคเหมืองข้อความ
Automatic Question Classification on Webboard Using Text Mining Techniques

ราชวิทย์ ทิพย์เสนา[1], ฉัตรเกล้า เจริญผล[2], แกมกาญจน์ สมประเสริฐศรี[3]
Rachawit Tipsena[1], Chatklaw Jareanpon[2], Gamgarn Somprasertsri[3]
Received: 10 September 2013; Accepted: 12 December 2013

บทคัดย่อ
กระดานสนทนาเป็นสื่อในการแลกเปลี่ยนความคิดเห็นหรือซักถามข้อสงสัยที่ได้รับความนิยมและแพร่หลายในปัจจุบัน การนำกระดานสนทนามาใช้ในหน่วยงานหรือองค์กร เป็นการเพิ่มช่องทางการติดต่อสื่อสารอีกช่องทางหนึ่งที่ก่อให้เกิดประโยชน์และประสิทธิภาพในการให้บริการของหน่วยงาน แต่คำถามบนกระดานสนทนาที่ยังไม่มีการจัดหมวดหมู่อาจส่งผลกระทบต่อการตอบคำถาม ทำให้ผู้ตอบตอบคำถามไม่ตรงประเด็นหรือตอบคำถามไม่ได้เนื่องจากบางคำถามนั้นไม่ตรงกับหน่วยงานของตนเอง งานวิจัยนี้จึงนำเสนอการจัดกลุ่มคำถามอัตโนมัติบนกระดานสนทนาโดยใช้เทคนิคเหมืองข้อความ ซึ่งได้ทำการศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกข้อความของ 3 เทคนิควิธี คือ เทคนิคการหาเพื่อนบ้านใกล้ที่สุด เทคนิคต้นไม้ตัดสินใจ และเทคนิคการเรียนรู้เบย์อย่างง่าย ผลการเปรียบเทียบประสิทธิภาพแสดงให้เห็นว่า เทคนิคการหาเพื่อนบ้านใกล้ที่สุดให้ประสิทธิภาพในการจำแนกดีที่สุด โดยค่าความถูกต้องเท่ากับ 0.89 ค่าความเที่ยง เท่ากับ 0.9 ค่าความระลึก เท่ากับ 0.89 และค่า F-Measure เท่ากับ 0.892

คำสำคัญ: การจำแนกข้อความ กระดานสนทนา เหมืองข้อความ เทคนิคการหาเพื่อนบ้านใกล้ที่สุด เทคนิคต้นไม้ตัดสินใจ เทคนิคการเรียนรู้เบย์อย่างง่าย

Abstract
People use webbaords as a popular method to share their comments, answers, and questions. Webboards are usually used in large orginazations and are for useful and effective communication. However, unclassified questions might affect an answer such as no answer or no correct answer. This research proposed an automatic webboard classification system using text mining which compares three techniques: k-Nearest Neighbor, Decision Tree and Naïve Bayes. It was found that k-Nearest Neighbor is the most effective techniques at 0.89 with Precision at 0.9, Recall at 0.89, and F-Measure at 0.892.

Keywords: text classification, webboard, text mining, k-nearest neighbor technique, decision tree technique, naïve bayes technique

บทนำ
กระดานสนทนา (Webboard) เป็นเว็บไซต์ที่ใช้สำหรับเป็นสื่อกลางในการแลกเปลี่ยนบทสนทนาหรือความคิดเห็นต่าง ๆ ผ่านระบบอินเทอร์เน็ตที่ได้รับความนิยมและแพร่หลายในปัจจุบัน ผู้ใช้งานสามารถที่จะตั้งหัวข้อกระทู้เพื่อแลกเปลี่ยนประเด็นความเห็นได้อย่างอิสระตามหัวข้อที่ตนเองสนใจ กระดานสนทนามีการพัฒนาเพื่อจะนำไปใช้งานในรูปแบบที่หลากหลายแตกต่างกันออกไป เช่น เป็นช่องทางในการติดต่อ ประกาศข่าวสาร หรือสอบถามข้อมูล ก่อให้เกิดเป็นเครือข่ายสังคมผ่านอินเทอร์เน็ต (Social Network) ซึ่งเป็นแหล่งค้นคว้าข้อมูลที่สำคัญอีกรูปแบบหนึ่งในยุคปัจจุบัน อีกทั้งกระดานสนทนายังง่ายต่อการใช้งาน ผู้ใช้งานสามารถเรียนรู้และทำความเข้าใจได้ด้วยตนเอง จึงช่วยให้ผู้ใช้งานเกิดความสะดวกสบาย ช่วยลดขั้นตอนไม่ว่าจะเป็นเวลาหรือสถานที่ในการติดต่อสื่อสารได้เป็นอย่างดีการนำกระดานสนทนามาใช้ในหน่วยงานหรือองค์กร ถือได้ว่าเป็นอีกช่องทางหนึ่งในการ

[1] *นิสิตปริญญาโท, [2]อาจารย์, [3]ผู้ช่วยศาสตราจารย์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม จังหวัดมหาสารคาม 44150*
[1] *Master degree student, [2]Lecturer, [3]Assist. Prof. Faculty of Informatics, Mahasarakham University, Mahasarakham 44150, Thailand*

ติดต่อสื่อสาร ที่ก่อให้เกิดประโยชน์และเพิ่มประสิทธิภาพในการให้บริการของหน่วยงาน ความหลากหลายของข้อความจึงเป็นปัญหาในการตีความหมายที่ต้องพบเจอ โดยเฉพาะข้อความที่ไม่มีการจำแนกหมวดหมู่อย่างชัดเจน อาจส่งผลกระทบต่อการตอบคำถามหรือการให้ข้อมูลของหน่วยงานที่ถูกต้อง อย่างไรก็ตามแม้จะมีการจำแนกหมวดหมู่คำถามบนกระดานสนทนาอย่างชัดเจน แต่ในบางครั้งผู้ตั้งคำถามอาจตั้งคำถามผิดหมวดหมู่ เนื่องจากผู้ตั้งคำถามเป็นผู้เลือกหมวดหมู่เอง หมวดหมู่ที่เลือกอาจไม่สอดคล้องกับคำถาม ส่งผลต่อการตอบคำถามและผู้ตั้งคำถามได้รับคำตอบที่ไม่ชัดเจน เช่นเดียวกัน การจำแนกคำถามแบบอัตโนมัติบนกระดานสนทนาด้วยเทคนิคการทำเหมืองข้อความจึงเป็นแนวทางในการแก้ปัญหาการตั้งคำถามที่ผิดหมวดหมู่ได้เป็นอย่างดี

งานวิจัยนี้ได้ทำการศึกษาข้อมูลการใช้งานกระดานสนทนา โดยลักษณะของกระดานสนทนาจะอยู่ในรูปแบบของการถามตอบประเด็นปัญหาหรือการซักถามข้อสงสัยต่าง ๆ เกี่ยวกับการให้บริการของหน่วยงาน จากข้อมูลการใช้งานพบว่า กระดานสนทนายังไม่มีการจำแนกหมวดหมู่ของคำถาม ซึ่งคำถามมีทั้งส่วนที่เกี่ยวข้องและไม่เกี่ยวข้องกับหน่วยงาน จึงเกิดปัญหาในการตอบคำถามของผู้ตอบที่ไม่สามารถตอบคำถามให้ชัดเจนและตรงประเด็นได้ ซึ่งจากงานวิจัยที่เกี่ยวข้องแสดงให้เห็นถึงวิธีการสร้างดัชนีเอกสารที่เหมาะสมด้วยวิธี TFIDF-Weighting และเทคนิควิธีที่นำมาใช้ทดสอบที่แตกต่างกัน ซึ่งการกำหนดค่า k มีผลต่อค่าความถูกต้องในการจำแนกหมวดหมู่ อีกทั้งการจำแนกหมวดหมู่ข้อความแบบอัตโนมัติยังช่วยแก้ปัญหาในการจำแนกได้

ด้วยเหตุผลที่กล่าวมาข้างต้น งานวิจัยนี้จึงนำเทคนิคในการจำแนกข้อความ (Text Classification) โดยเป็นกระบวนการในการทำเหมืองข้อความ (Text Mining) มาช่วยในการจำแนกหมวดหมู่ของคำถามแบบอัตโนมัติบนกระดานสนทนา โดยจะพิจารณาจากคุณลักษณะของคำในแต่ละคำถามในการจำแนกหมวดหมู่ เพื่อแยกประเภทและความแตกต่างของคำถาม ทำให้ผู้ตอบสามารถตอบคำถามได้ตรงกับการให้บริการของหน่วยงาน และผู้ตั้งคำถามได้รับคำตอบที่มีความถูกต้องชัดเจน

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1. เหมืองข้อความ

การทำเหมืองข้อความ[1] เป็นกระบวนการค้นหาความรู้หรือข้อเท็จจริงที่แฝงอยู่ในชุดข้อความ การทำเหมืองข้อความมักจะนำไปใช้กับข้อมูลประเภทไม่มีโครงสร้าง (Unstructured Data) เช่น ข้อความอีเมล์ ข้อความบนเว็บบอร์ด หรือข้อมูลประเภทกึ่งโครงสร้าง (Semi-Structured Data) เช่น ข้อความรูปแบบ XML หรือ HTML เหมืองข้อความจึงเป็นเครื่องมือในการจัดการข้อมูลเหล่านี้ได้เป็นอย่างดี แต่การทำเหมืองข้อความมีความแตกต่างจากการทำเหมืองข้อมูลในด้านเครื่องมือโดยเครื่องมือการสร้างเหมืองข้อมูลจะออกแบบมาสำหรับข้อมูลที่มีโครงสร้าง ส่วนเครื่องมือการทำเหมืองข้อความจะออกแบบมาสำหรับข้อมูลที่ไม่มีโครงสร้าง

ปัจจุบันเหมืองข้อความได้รับความนิยมและนำมาใช้ประโยชน์ทางด้านการค้นหาข้อมูลที่แฝงอยู่ในชุดข้อความที่มีจำนวนมหาศาล ทำให้ได้ข้อมูลที่เป็นประโยชน์และช่วยสนับสนุนการตัดสินใจในด้านต่าง ๆ เช่น การทำเหมืองข้อความเพื่อวิเคราะห์ข้อมูลการให้บริการ เพื่อนำผลการวิเคราะห์มาสนับสนุนการวางแผนการตลาด เป็นต้น[2,3]

กระบวนการทำเหมืองข้อความ มีกระบวนการคล้ายคลึงกับกระบวนการค้นหาความรู้จากฐานข้อมูล ทั้งนี้หากผลจากการวิเคราะห์ในแต่ละขั้นตอนมีความถูกต้องหรือความน่าเชื่อถือต่ำเกินไป จะต้องกลับไปขั้นตอนที่ต่ำกว่า หรือทำการเลือกข้อมูลมาใหม่เพื่อทำให้กระบวนการทำเหมืองข้อความมีคุณภาพ โดยสามารถแบ่งกระบวนการทำงานออกเป็น 5 ขั้นตอนสำคัญ[4] ดังภาพที่ 1 มีกระบวนการดังนี้

1. การเลือกข้อมูล (Selection) เป็นการระบุถึงแหล่งข้อมูลที่จะนำมาใช้ในการทำเหมืองข้อความ รวมถึงการนำข้อมูลที่ต้องการออกมาจากฐานข้อมูล เพื่อทำการพิจารณาในเบื้องต้นตามขอบเขตที่ต้องการทำการศึกษา

2. การเตรียมข้อมูล (Preprocessing) เป็นกระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูลที่จะนำมาใช้วิเคราะห์ว่ามีความถูกต้อง โดยการนำข้อมูลที่ไม่ถูกต้องออกหรือเป็นขั้นตอนที่อาจต้องแก้ไขข้อมูลก่อนนำไปใช้งาน

3. การจัดทำดัชนีข้อมูล (Indexing) เป็นการจัดข้อมูลให้เหมาะสมและตรงกับรูปแบบที่จะประมวลผลต่อไป เช่น การตัดบางคอลัมน์ที่ไม่จำเป็นออก

4. การทำเหมืองข้อมูล (Data mining) เป็นขั้นตอนประมวลผล โดยใช้อัลกอริทึมต่าง ๆ เพื่อแก้ไขปัญหาหรือหารูปแบบของข้อมูล การจัดหมวดหมู่ (Classification) การค้นหากฎความสัมพันธ์ (Association Rule Discovery) การแบ่งกลุ่มข้อมูล (Data Clustering) และการสร้างจินตทัศน์ (Visualization)

5. การแปลผลและการประเมินผล (Interpretation/Evaluation) เป็นขั้นตอนการแปลความหมาย การตีความและการประเมินผลลัพธ์ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ ซึ่งควรมีการนำเสนอผลการวิเคราะห์ในรูปแบบที่ผู้ใช้งานสามารถเข้าใจได้ง่าย

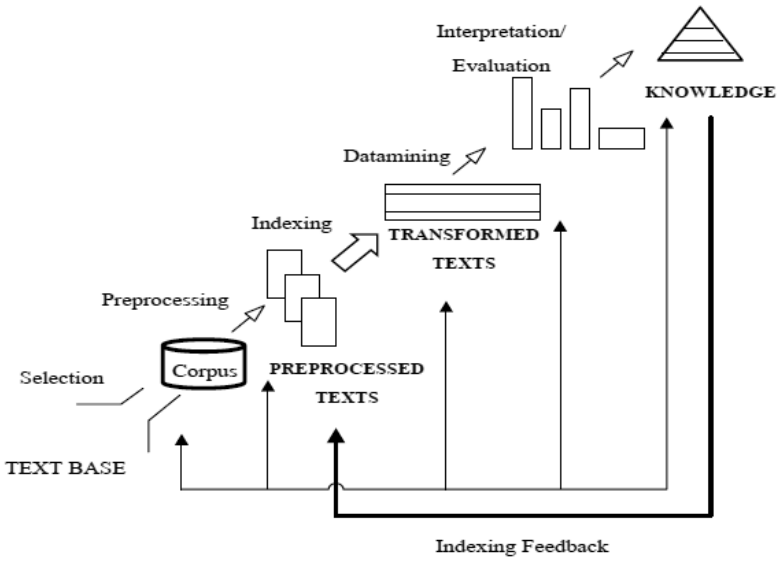

Figure 1 Process of Text mining[4]

2. การประมวลผลข้อความ

การประมวลผลข้อความ (Text Preprocessing) เป็นการจำแนกเอกสารภาษาไทยอัตโนมัติ มีขั้นตอนการทำงาน คือ ขั้นตอนแรกจะทำการสกัดคุณลักษณะด้วยการตัดคำเพื่อให้ได้คุณลักษณะจากเอกสารออกมา จากนั้นทำการกำจัดคำหยุดและทำรากศัพท์จากฐานข้อมูลภาษาไทยที่กำหนดขึ้น หลังจากนั้นทำการให้ค่าน้ำหนักดัชนีของคำในเอกสาร (Term Weighting) แล้วทำการลดขนาดคุณลักษณะเพื่อมิติของเอกสารลดลง จากนั้นให้ทำการเรียนรู้แบบมีผู้สอน Supervised Learning)[5] มีรายละเอียดของแต่ละขั้นตอน ดังนี้

1. การตัดคำ (Word Segmentation) เป็นขั้นตอนการประมวลผลการจำแนกหมวดหมู่ข้อความ เพื่อให้การจำแนกมีประสิทธิภาพ สำหรับการตัดคำในภาษาไทย ยังพบปัญหาในการตัดคำเนื่องจากลักษณะของภาษาไทยมีการเขียนติดต่อกันแบบไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำที่ชัดเจน ไม่มีช่องว่างคั่นที่แสดงให้เห็นถึงขอบเขตของแต่ละคำ จึงมีผู้คิดค้นวิธีการตัดคำภาษาไทยโดยมีวิธีการ 3 วิธีหลัก[5,6] ดังนี้

หลักการตัดคำโดยใช้กฎ (Rule-Based Approach) เป็นการตัดคำโดยตรวจสอบจากกฎเกณฑ์ทางอักขระวิธีโดยอาศัยหลักไวยากรณ์ภาษาไทย ซึ่งเริ่มจากการตัดพยางค์ เนื่องจากพยางค์มีรูปแบบที่แน่นอนมากกว่าคำ จากนั้นจึงนำพยางค์มาเป็นเกณฑ์ในการกำหนดขอบเขตคำ ข้อดีคือ การทำงานมีความรวดเร็วและใช้ทรัพยากรในการประมวลผลน้อย แต่มีข้อจำกัดคือ ผลของการตัดคำอาจได้เป็นกลุ่มคำที่สามารถตัดคำออกไปได้อีก

หลักการตัดคำโดยใช้พจนานุกรม (Dictionary-Based Approach) ใช้การเปรียบเทียบคำกับคำที่จัดเก็บในพจนานุกรมร่วมกับการใช้กฎในการตัดคำ คำทั้งหมดจะถูกจัดเก็บไว้ในพจนานุกรม แล้วนำข้อความที่ป้อนเข้ามาไปเปรียบเทียบสายอักขระกับคำในพจนานุกรม ปัญหาของวิธีการนี้ คือ ไม่สามารถจัดเก็บคำทั้งหมดลงพจนานุกรมได้ เนื่องจากมีคำใหม่ ๆ เกิดขึ้น ดังนั้นความถูกต้องจะขึ้นอยู่กับปริมาณของคำในพจนานุกรม และต้องเพิ่มพื้นที่ในการจัดเก็บตามปริมาณคำที่เพิ่มมากขึ้น

หลักการตัดคำโดยใช้คลังข้อมูล (Corpus-Based Approach) เป็นการนำหลักทางสถิติและกลไกการเรียนรู้มาใช้ในการประมวลผลทางภาษา โดยอาศัย 2 วิธี คือ วิธีอาศัยค่าความน่าจะเป็น โดยใช้แบบจำลองไตรแกรมกำกับหน้าที่ของคำ ในการหารูปแบบของการตัดคำและลำดับหมวดหมู่ วิธีการนี้จะต้องเตรียมคลังข้อมูลที่มีการตัดคำและการกำกับหน้าที่ของคำไว้ก่อนล่วงหน้า และวิธีอาศัยคุณลักษณะของคำ จะช่วยแก้ไขความผิดพลาดของการตัดคำที่อาศัยค่าความน่าจะเป็น โดยนำคุณลักษณะคำที่มีความกำกวมมาช่วยเลือกการตัดคำที่ถูกต้อง ให้ได้จำนวนคำที่ไม่พบในพจนานุกรมน้อยที่สุด โดยการสร้างแบบจำลองกลไกการเรียนรู้

งานวิจัยนี้ใช้โปรแกรมเล็กซ์โต (Thai Lexeme Tokenizer : LexTo)[7] ในการตัดคำข้อความภาษาไทย เป็นหลักการตัดคำโดยใช้พจนานุกรม เทียบคำที่ยาวที่สุดที่พบในพจนานุกรม (Longest Matching) หลักการทำงาน คือ เมื่อมีการส่งข้อความเข้ามา โปรแกรมจะอ่านคำทั้งหมดแล้วเก็บไว้ในโปรแกรม แล้วจะนำคำเหล่านั้นไปเทียบกับคำในพจนานุกรมของโปรแกรม ตัวอย่างการตัดคำด้วยโปรแกรม

เล็กซ์โต Figure 2 และคำศัพท์ที่ได้หลังจากการตัดคำ Table 1

Figure 2 LexTo software[7]

Table 1 Example for word segmentation

นักศึกษา	พัน	สภาพ	เนื่องจาก	เกรดเฉลี่ย
ไม่ถึง	2	อยาก	กลับ	เข้ามา
ใหม่	ใน	คณะ	เดิม	แต่
ไม่ทราบ	ข้อมูล	การ	รับสมัคร	ว่า
ต้อง	ทำ	อย่างไรบ้าง	?	รบกวน
ผู้อำนวยการ	ช่วย	แนะนำ	ด้วย	ค่ะ

2. การกำจัดคำหยุด (Stop-Word Removal)[8,9] เป็นการนำคำที่ไม่มีนัยสำคัญออก โดยที่ความหมายของคำหรือข้อความไม่เปลี่ยนแปลง คำหยุดจะปรากฏอยู่ในข้อความทุกข้อความ มีความถี่สูง ถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการจำแนกหมวดหมู่ การกำจัดคำหยุดเป็นกระบวนการที่จะช่วยให้ขนาดของดัชนีลดลง และยังลดขนาดพื้นที่และเวลาในการประมวลผล ประเภทของคำที่เป็นคำหยุดในภาษาไทย ได้แก่ คำบุพบท คำสันธาน คำสรรพนาม คำวิเศษณ์ และคำอุทาน ตัวอย่างคำหยุด Table 2

Table 2 Thai language stops words

คำบุพบท	คำสันธาน	คำสรรพนาม	คำวิเศษณ์	คำอุทาน
กับ	กว่า	กระผม	ใกล้	555
ของ	คือ	ข้าพเจ้า	ครับ	กำ
ซึ่ง	จึง	ฉัน	ง่าย	ขอโทษ
ด้วย	แต่	เธอ	ด้วย	ไง
โดย	ถ้า	นาย	ที่สุด	เยี่ยม

3. การหารากศัพท์ (Stem)[10] คือ รูปแบบคำที่ยังไม่เปลี่ยนรูป หรือยังไม่เติมคำอุปสรรค (Prefixes), คำปัจจัย (Suffixes) เป็นการหารูปแบบเดิมของคำหรือหาคำที่มีความหมายคล้ายกันมารวมเป็นคำเดียวกัน แต่ยังมีข้อจำกัด คือ ยังไม่มีอัลกอริทึมสำหรับการหารากศัพท์ เนื่องจากไวยากรณ์ภาษาไทยมีความซับซ้อน ดังนั้นการหารากศัพท์ภาษาไทยสามารถใช้วิธีรวมคำศัพท์ที่มีความหมายคล้ายกันให้เป็นรากศัพท์ แล้วจัดเก็บลงคลังคำ เพื่อใช้เปรียบเทียบคำในการหารากศัพท์ โดยจะต้องให้ผู้เชี่ยวชาญด้านภาษาไทยเป็นผู้จัดทำคลังคำสำหรับรากศัพท์ภาษาไทย ตัวอย่างรากศัพท์ Table 3[11]

Table 3 Thai language stemming

รากศัพท์	คำศัพท์
ปริญญา	ปริญญากิตติมศักดิ์, ปริญญาตรี, ปริญญาโท
เกษตร	เกษตรกร, เกษตรกรรม, เกษตรศาสตร์
คณะ	คณะกรรมการ, คณะกรรมาธิการ, คณะทูต
นาม	นามบัตร, นามปากกา, นามแฝง, นามธรรม
ฤกษ์	ฤกษ์ดี, ฤกษ์บน, ฤกษ์พานาที, ฤกษ์งามยามดี

4. การสร้างดัชนีคำสำคัญ (Indexing)[6,12,13] เป็นกระบวนการแปลงเอกสารที่เป็นภาษาธรรมชาติ ให้คอมพิวเตอร์สามารถเข้าใจและประมวลผลได้ การสร้างดัชนีเป็นการสร้างตัวแทนเอกสาร (Document Representation) ตัวแทนเอกสารโดยทั่วไปจะอยู่ในรูปแบบของเวกเตอร์ค่าน้ำหนักคำ คือ การคำนวณหาค่าที่จะมาใช้เป็นค่าคุณลักษณะของเอกสาร หรือการหาค่าน้ำหนัก (Term Weighting) เช่น คำเดี่ยว (Single Word), รากศัพท์ (Stem), วลี (Phrase), ชุดลำดับ (N-Gram), ประโยค (Sentence) เป็นต้น

งานวิจัยนี้ใช้การสร้างดัชนีเอกสารโดยการหาค่าน้ำหนักรูปแบบคำเดี่ยว โดยใช้วิธี TFIDF-Weighting (Term Frequency Inverse Document Frequency) ดังสมการที่ 1[5]

$$tfidf_{ik} = (1+\log(tf_{ik})) \times \log(\frac{N}{n_i}) \qquad (1)$$

โดยที่

N คือ จำนวนเอกสารทั้งหมด

n_i คือ จำนวนเอกสารใน N ที่มีคำ tf_{ik} ปรากฏอยู่

5. การเลือกคุณลักษณะ (Feature Selection)[14,15] เป็นการสร้างคุณลักษณะใหม่จากคุณลักษณะเดิม เนื่องจากจำนวนคุณลักษณะมีผลต่อประสิทธิภาพการจำแนกหมวดหมู่ข้อความ การเลือกคุณลักษณะที่นิยม ได้แก่ ค่าความถี่เอกสาร (DF: Document Frequency), ค่าสารสนเทศ (IG: Information Gain), ค่าสถิติไคสแควร์ (Chi-Square)

3. การจำแนกหมวดหมู่ข้อความ

การจำแนกหมวดหมู่ข้อความเป็นการแยกข้อความซึ่งประกอบด้วยภาษาธรรมชาติ ให้อยู่ภายใต้หมวดหมู่ที่กำหนดไว้ก่อนโดยอาศัยตามตัวจำแนก ในการกำหนดหมวดหมู่จะพิจารณาจากความคล้ายคลึงระหว่างชุดข้อมูลเรียนรู้และชุดข้อมูลจริง และอาจจะพิจารณาจากลักษณะความหมายของคำ

งานวิจัยนี้ได้นำเทคนิคการหาเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbor) เทคนิคต้นไม้ตัดสินใจ (Decision Tree) และเทคนิคการเรียนรู้เบย์อย่างง่าย (Naïve Bayes) เป็นอัลกอริทึมในการทดสอบการจำแนกข้อความ

4. งานวิจัยที่เกี่ยวข้อง

Choochart Haruechaiyasak และคณะ[15] ได้นำเสนอการให้ค่าน้ำหนักดัชนีเอกสารด้วยวิธี TFIDF-Weighting โดยทดลองกับกลุ่มตัวอย่างหนังสือพิมพ์ไทยรัฐจากเว็บไซต์ โดยใช้การตัดคำด้วยคำเดี่ยว (Single Word) ผลการทดลองพบว่าเมื่อใช้อัลกอริทึม Support Vector Machines ให้ประสิทธิภาพออกมาสูงสุด และ Naïve Bayes, Decision Tree รองลงมาตามลำดับ TFIDF-Weighting เป็นวิธีที่นิยมใช้ในการสร้างแบบจำลองเอกสาร เนื่องจากมีประสิทธิภาพในการจำแนกดี ให้ความแม่นยำสูง

Shiqun Yin และคณะ[16] ศึกษาเกี่ยวกับการทำเหมืองข้อมูลข้อความบนเว็บไซต์ โดยสกัดเนื้อหาเพื่อการทำเหมืองข้อมูล การจัดหมวดหมู่ข้อความในรูปแบบการแบ่งประเภทข้อความออกเป็นชุดของข้อความ โดยใช้อัลกอรึทึม Vector Space Model (VSM) ในการจัดหมวดหมู่ข้อความแบ่ง k=10 ค่าความถูกต้องในการจัดหมวดหมู่ได้เท่ากับ 81% ใช้เวลาประมาณ 4 วินาที จึงทำให้ทราบถึง กระบวนการในการทำเหมืองข้อความและการแบ่งจำนวน k ที่ให้ค่าความถูกต้องในการจัดหมวดหมู่ที่ดี

Huan Huang และคณะ[17] ได้ศึกษาการจัดหมวดหมู่ข้อความสั้นแบบอัตโนมัติ โดยการเชื่อมโยงเนื้อหาของข้อความให้ตรงกับประเภทหมวดหมู่ข้อความโดยอัตโนมัติ โดยใช้อัลกอริทึม Support Vector Machine (SVM) และ k-Nearest Neighbor (k-NN) ใช้คำในการจัดหมวดหมู่ คุณสมบัติจึงขึ้นอยู่กับความถี่ของคำที่เพิ่มสูงขึ้น คำหรือข้อความทั่วไปไม่สามารถเป็นตัวแทนของข้อความที่ดีได้ งานวิจัยได้แสดงให้เห็นถึงผลการจัดหมวดหมู่ข้อความแบบอัตโนมัติ โดยแก้ปัญหาคำหรือข้อความทั่วไปที่ไม่สามารถเป็นตัวแทนของข้อความที่ดีได้ จึงใช้การสกัดคำจากหัวข้อในการแก้ปัญหา

Jiabin Deng และคณะ[18] {Deng, 2010 #6} {Deng, 2010 #6}{Deng, 2010 #6}นำเสนอวิธีการปรับปรุงการจัดกลุ่มข้อความโดยใช้อัลกอริทึม C-Means และการแก้ไขการคำนวณระยะห่าง เนื่องจากการจัดกลุ่มด้วยอัลกอริทึม C-Means มีความเที่ยงตรงน้อย งานวิจัยจึงแนะนำตัวอย่างที่ให้ค่าความถูกต้องสูง โดยการปรับ Radius และ Weight วิธีการแก้ไขแสดงให้เห็นว่าผลลัพธ์ของอัลกอริทึมในการจัดกลุ่มข้อความมีประสิทธิภาพมากขึ้น ผลการวิจัยแสดงให้เห็นว่าก่อนการปรับ Radius และ Weight จะได้ค่าความถูกต้องเท่ากับ 86.52% เวลาเท่ากับ 10.32 วินาที หลังทำการปรับได้ค่าความถูกต้องเท่ากับ 88.12% เวลาเท่ากับ 4.21 วินาที ซึ่งการปรับค่า Radius และ Weight มีผลทำให้ได้ค่าความถูกต้องและมิติของเวลามีประสิทธิภาพดีขึ้น

การดำเนินงานวิจัย

การดำเนินงานวิจัยจะนำเสนอกระบวนการทำเหมืองข้อความเพื่อจำแนกกลุ่มข้อความและเปรียบเทียบประสิทธิภาพการจำแนกของ 3 เทคนิควิธี แบ่งขั้นตอนการทำงาน ดังนี้

1. การเก็บรวบรวมข้อมูล

งานวิจัยนี้เก็บรวบรวมข้อมูลจากกระดานสนทนา กองบริการการศึกษา มหาวิทยาลัยมหาสารคาม ซึ่งเป็นเว็บไซต์ของหน่วยงานทางด้านการรับเข้าศึกษา ระดับปริญญาตรี โดยเก็บข้อมูลคำถาม ตั้งแต่เดือนมิถุนายน 2554 ถึงเดือนตุลาคม 2556 จำนวน 1,165 เรคคอร์ด มีรายละเอียดและขอบเขตข้อมูล Table 4

Table 4 Attribute of questions

No	Filed	Description	Type
1	topicID	รหัสคำถาม	Int(5) PK
2	Title	คำถาม	Varchar(255)
3	Detail	รายละเอียด	Text
4	postName	ชื่อผู้ตั้งคำถาม	Varchar(50)
5	contact	ข้อมูลติดต่อ	Varchar(50)
6	postDate	วัน เวลาถาม	Date time

3. การประมวลผลข้อความ

เนื่องจากข้อความที่ได้มายังอยู่ในรูปแบบภาษาธรรมชาติ จึงต้องแปลงข้อความให้อยู่ในรูปแบบที่คอมพิวเตอร์

สามารถเรียนรู้ได้ เพื่อสร้างตัวแทนเนื้อหาของเอกสาร โดยกระบวนการประมวลผลข้อความสำหรับงานวิจัยนี้มีขั้นตอน Figure 3

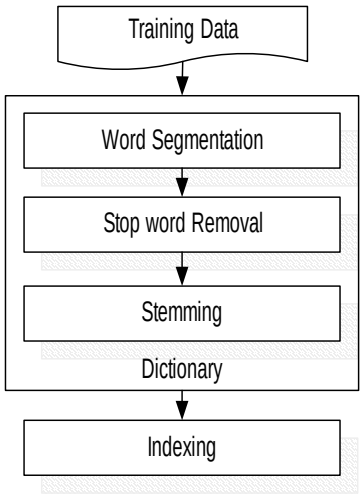

Figure 3 Text preprocessing

ขั้นตอนการตัดคำ การกำจัดคำหยุด และการหารากศัพท์ในภาษาไทย เป็นกระบวนการที่ต้องอาศัยพจนานุกรมคำศัพท์มาช่วยในกระบวนการเหล่านี้ โดยงานวิจัยใช้พจนานุกรมคำศัพท์เล็กซิตรอน (Lexitron)[11] เพื่ออ้างอิงคำศัพท์ที่ถูกต้อง

งานวิจัยนี้แบ่งคลาสข้อมูลตามลักษณะของคำถามที่พบบนกระดานสนทนา มีผู้เชี่ยวชาญในหน่วยงานเป็นผู้เลือกคำสำคัญของแต่ละคลาสข้อมูล โดยแบ่งคลาสข้อมูลออกเป็น 4 คลาส ได้แก่ คลาส Admissions มี 480 คำถาม, คลาส Fee มี 212 คำถาม, คลาส Register มี 338 คำถาม, คลาส Other มี 135 คำถาม มีคำสำคัญรวมทั้งหมดจำนวน 102 คำ และจัดเก็บคำสำคัญในลักษณะของถุงคำ Table 5 Table 6 Table 7 และ Table 8

Table 5 Words of Admissions class

ระบบ	เข้า	สำรอง	ปวช.
พิเศษ	สัมภาษณ์	ติด	ระเบียบการ
สมัคร	ต่อ	โควตา	คัดเลือก
คณะ	สาขา	Admissions	กลาง
รับ	แอด	ตัดสิทธิ์	เลือก
รอบ	ประกาศ	เรียก	ทั่วประเทศ
ตรง	ปวส.	ต่อเนื่อง	เกณฑ์

Table 6 Words of Fee class

จ่าย	เท่าไร	กู้
ค่า	ใบเสร็จ	ค่าปรับ
เงิน	ค่าใช้จ่าย	ค่าเล่าเรียน
ค่าธรรมเนียม	ธนาคาร	หนี้
ชำระ	ค้าง	เบิก
ตัง	สลิป	ยอด

Table 7 Words of Register class

นิสิต	จบ	หลักสูตร	ผลการเรียน
ย้าย	สภาพ	วิชา	มอบตัว
รายงานตัว	เทียบ	ลาพัก	คำร้อง
ลงทะเบียน	ทะเบียน	รับรอง	วุฒิ
หน่วยกิต	เกรด	ลาออก	รีไทร์
รหัส	ซิ่ว	ภาคเรียน	ปริญญา
เทอม	สิทธิ์	ทรานสคริป	โอน

Table 8 Words of Other class

มมส	มหาลัย	แต่งกาย	ปรึกษา
สอบถาม	ตอบ	ติดต่อ	สั่งซื้อ
หอพัก	เว็บ	สถาบัน	จดหมาย
มหาวิทยาลัย	กยศ	วิชาชีพ	เครือข่าย
มหาสารคาม	จอง	สถานที่	คุณภาพ
ชุด	บริการ	ข้าราชการ	โรงเรียน
กอง	ฉุกเฉิน	ประกอบการ	ร้องเรียน

การสร้างดัชนีคำสำคัญ TFIDF-Weighting อยู่ในรูปแบบของเวกเตอร์เอกสาร (Word Vector) โดยค่าน้ำหนักของคำหาได้จากการเปรียบเทียบคำถามกับคำสำคัญในถุงคำ แล้วนับจำนวนคำสำคัญที่พบในคำถาม เพื่อนำความถี่มาคำนวณหาค่าน้ำหนัก Table 9

Table 9 TFIDF-Weighting word vector

Question	w_1	w_2	w_3	w_4	…	w_{102}	class
Q1	0	0.2	0	0.1	…	0	A
Q2	0.1	0.1	0.3	0	…	0	A
Q3	0	0.1	0	0	…	0.1	F
Q4	0.2	0	0	0	…	0	R
Q5	0.1	0	0.1	0	…	0	R
Q6	0	0	0	0	…	0	O
…	0	0	0	0	…	0	F
Q_{1165}	0	0	0.2	0	…	0	O

3. สร้างแบบจำลองจำแนกกลุ่มข้อความ

การแบ่งข้อมูลทดสอบแบ่งโดยใช้วิธีการประเมินผล K-fold Cross Validation กำหนด k-fold เป็น 3 กลุ่ม ได้แก่ k=10, k=15, k=20 สำหรับการทดสอบข้อมูลจะใช้โปรแกรม Weka (Waikato Environment for Knowledge Analysis) สำหรับเทคนิคการหาเพื่อนบ้านใกล้ที่สุด ทดสอบ k ได้แก่ 3NN, 5NN, 7NN, 9NN แล้วเลือก k ที่ให้ผลลัพธ์ดีที่สุด จากนั้นจะหาค่าเฉลี่ยการวัดประสิทธิภาพของแต่ละเทคนิควิธี นำแบบจำลองของเทคนิควิธีที่ให้ประสิทธิภาพดีที่สุดมาพัฒนาระบบจำแนกคำถามอัตโนมัติ โดยมีกระบวนการ Figure 4

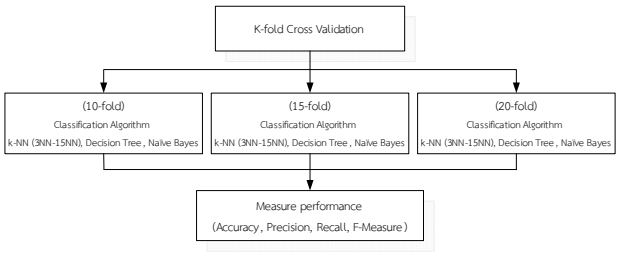

Figure 4 Process of Classification model

4. ประเมินประสิทธิภาพการทำงาน

การประเมินประสิทธิภาพการทำงาน ใช้การวัดประสิทธิภาพโดยการหา ค่าความเที่ยง (Precision) ค่าความระลึก (Recall) ค่าความถูกต้อง (Accuracy) และค่า F-Measure ดังสมการที่ 2, 3, 4 และ 5 ตามลำดับ

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (3)$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad (4)$$

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (5)$$

โดยที่ TP คือ จำนวนข้อมูลที่ถูกดึงออกมาอย่างถูกต้อง
FP คือ จำนวนข้อมูลที่ผิดพลาดที่ถูกดึงออกมา
TN คือ จำนวนข้อมูลที่ถูกต้องแต่ไม่ถูกดึงออกมา
FN คือ จำนวนข้อมูลที่ผิดพลาดแต่ไม่ถูกดึงออกมา

5. ออกแบบการทำงานผ่านเว็บไซต์

นำแบบจำลองที่ได้จากการทำเหมืองข้อความมาประยุกต์ใช้งานบนกระดานสนทนาในการจำแนกหมวดหมู่คำถามอัตโนมัติ โดยมีขั้นตอนการพัฒนา Figure 5

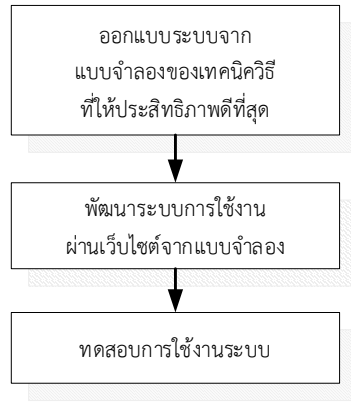

Figure 5 Process of System design

ขั้นตอนที่ 1 นำเทคนิควิธีที่ให้ประสิทธิภาพในการจำแนกดีที่สุดมาออกแบบเพื่อพัฒนาเป็นโปรแกรม โดยศึกษาหลักการทำงานของเทคนิควิธีเพื่อนำมาพัฒนาระบบจำแนกกลุ่มคำถามอัตโนมัติ

ขั้นตอนที่ 2 พัฒนาระบบการใช้งาน โดยระบบจะอยู่ในรูปแบบของเว็บไซต์แสดงผลผ่านเว็บบราวเซอร์ (Web Brower) ซึ่งพัฒนาด้วยภาษา PHP : Hypertext Preprocessor และระบบฐานข้อมูล MySQL โดยยึดกระบวนการจากแบบจำลองของเทคนิควิธีที่ดีที่สุด

ขั้นตอนที่ 3 ทดสอบการใช้งานระบบ โดยการป้อนคำถามใหม่ ระบบจะนำข้อความมาเปรียบเทียบกับคำสำคัญในถุงคำซึ่งเก็บอยู่ในรูปแบบของฐานข้อมูลแล้วนับจำนวนคำที่พบเพื่อนำความถี่ของคำไปประมวลผลตามแบบจำลองที่ได้จากเทคนิควิธีที่ดีที่สุด

ผลการดำเนินงาน

1. ผลการจำแนกและทดสอบประสิทธิภาพ

ผลการจำแนกกลุ่มข้อความจาก 3 เทคนิควิธีได้ผลลัพธ์ค่าความถูกต้อง ค่าความเที่ยง ค่าความระลึก และค่า F-Measure ของแต่ละเทคนิควิธี โดยเทคนิคการหาเพื่อนบ้านใกล้ที่สุดผลดังตารางที่ 10 เทคนิคต้นไม้ตัดสินใจและเทคนิคการเรียนรู้เบย์อย่างง่ายผล Table 11

Table 10 Result of k-Nearest Neighbor

k-fold	Nearest Neighbor	Accuracy	Precision	Recall	F-Measure
10-fold	3NN	88.58%	0.896	0.886	0.888
	5NN	86.69%	0.881	0.867	0.870
	7NN	85.57%	0.872	0.856	0.859
	9NN	84.12%	0.865	0.841	0.846
15-fold	3NN	89.01%	0.900	0.890	0.892
	5NN	87.46%	0.886	0.875	0.877
	7NN	85.23%	0.871	0.852	0.856
	9NN	84.72%	0.866	0.847	0.851
20-fold	3NN	88.32%	0.894	0.883	0.885
	5NN	86.69%	0.881	0.867	0.87
	7NN	85.92%	0.877	0.859	0.863
	9NN	83.94%	0.866	0.839	0.845

Table 11 Result of Decision Tree and Naïve Bayes

k-fold	Algorithm	Accuracy	Precision	Recall	F-Measure
10-fold	Decision Tree	86.09	0.881	0.861	0.865
	Naïve Bayes	86.86	0.898	0.869	0.876
15-fold	Decision Tree	87.03	0.889	0.870	0.875
	Naïve Bayes	87.38	0.901	0.874	0.881
20-fold	Decision Tree	86.43	0.879	0.864	0.867
	Naïve Bayes	87.21	0.900	0.872	0.879

จาก Table 10 และ Table 11 เมื่อเปรียบเทียบผลลัพธ์ค่าความถูกต้อง ค่าความเที่ยง ค่าความระลึก และค่า F-Measure จากทั้ง 3 เทคนิควิธีแสดงให้เห็นว่า เทคนิคการหาเพื่อนบ้านใกล้ที่สุดให้ประสิทธิภาพในการจำแนกดีที่สุด โดยจำนวน k-fold สำหรับการทดสอบข้อมูลที่ให้ประสิทธิภาพดีที่สุดเท่ากับ 15-fold และ k-NN เท่ากับ 3 ได้ค่าความถูกต้องเท่ากับ 89.01% ค่าความเที่ยงเท่ากับ 0.9 ค่าความระลึกเท่ากับ 0.89 และค่า F-Measure เท่ากับ 0.892

จากผลการทดสอบและเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มข้อความ จะเห็นได้ว่าการจำแนกด้วยเทคนิคการหาเพื่อนบ้านใกล้ที่สุดให้ประสิทธิภาพในการจำแนกดีที่สุด ผู้วิจัยจึงนำหลักการของเทคนิควิธีการหาเพื่อนบ้านใกล้ที่สุดมาสร้างแบบจำลองในการจำแนกกลุ่มคำถามอัตโนมัติ เพื่อให้สามารถใช้งานด้านการจำแนกกลุ่มคำถามบนกระดานสนทนาได้

2. ผลการพัฒนาระบบจำแนกกลุ่มคำถาม

จากการวิเคราะห์แบบจำลองการจำแนกกลุ่มคำถาม สามารถพัฒนาระบบในการจำแนกคำถามอัตโนมัติ ซึ่งอยู่ในรูปแบบของเว็บแอพพลิเคชั่น โดยกระบวนการจำแนกกลุ่มคำถามอัตโนมัติจะใช้หลักการคำนวณหาค่า Distance ของเทคนิควิธีการหาเพื่อนบ้านใกล้ที่สุดเนื่องจากให้ประสิทธิภาพในการจำแนกดีที่สุด โดย k = 3NN เพื่อหากลุ่มให้กับคำถามบนกระดานสนทนา ผลการพัฒนาระบบจำแนก Figure 6

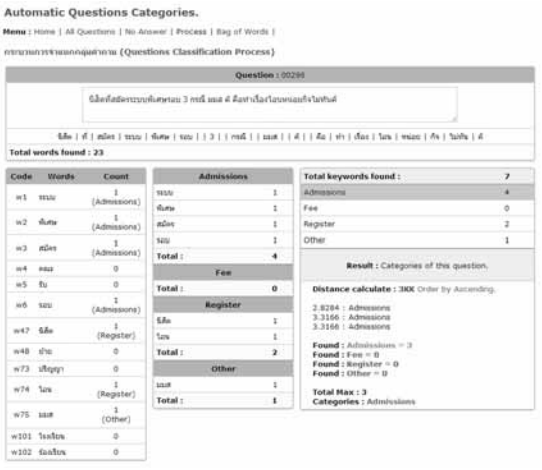

Figure 6 Result of Question classification

สรุปผล

ผลการจำแนกข้อความภาษาไทยที่ได้จากคำถามบนกระดานสนทนา โดยการเปรียบเทียบประสิทธิภาพการทำงานของอัลกอริทึม 3 เทคนิควิธี พบว่า เทคนิควิธีการหาเพื่อนบ้านใกล้ที่สุดให้ประสิทธิภาพดีที่สุด โดย k ที่ดีที่สุดเท่ากับ 3 และ K-fold Cross Validation เท่ากับ 15-fold ค่าความถูกต้องที่ได้เท่ากับ 89.01% ค่าความเที่ยงเท่ากับ 0.9 ค่าความระลึกเท่ากับ 0.89 และค่า F-Measure เท่ากับ 0.892 แสดงให้เห็นว่าประสิทธิภาพการจำแนกข้อความขึ้นอยู่กับเทคนิควิธีที่นำมาใช้จำแนก และการกำหนด K-fold ที่เหมาะสม หากเป็นเทคนิควิธีการหาเพื่อนบ้านใกล้ที่สุด ประสิทธิภาพยังขึ้นอยู่กับการทดสอบค่า k อีกด้วย สำหรับการวิจัยสามารถเปลี่ยนเทคนิควิธีในการจำแนกข้อความเพื่อทดสอบประสิทธิภาพการจำแนกได้ และสามารถนำไปพัฒนาแบบจำลองทางด้านเหมืองข้อความเพื่อประยุกต์ใช้งานทางด้านการจำแนกข้อความได้เป็นอย่างดี

งานวิจัยนี้ได้ประยุกต์ใช้งานกระบวนการจำแนก โดยการพัฒนาระบบจำแนกกลุ่มข้อความอัตโนมัติ จากการนำกระบวนการของเทคนิควิธีที่ให้ประสิทธิภาพดีที่สุด คือ เทคนิคการหาเพื่อนบ้านใกล้ที่สุด มาพัฒนาเป็นระบบจำแนกกลุ่มคำถาม ซึ่งการคำนวนหาค่า Distance และการกำหนดค่า k ของ k-NN จะช่วยในการจำแนกกลุ่มคำถามได้เป็นอย่างดี

ข้อเสนอแนะ

1. การกำจัดคำหยุด ต้องคำนึงถึงความสำคัญของคำ เนื่องจากคำหยุดบางคำอาจจะนำมาใช้เป็นคำสำคัญได้ ขึ้นอยู่กับลักษณะของคำที่จะนำมาใช้ในการประมวลผล

2. พิจารณาว่าในการประมวลผลคำมีความจำเป็นต้องหารากศัพท์หรือไม่ ซึ่งคำในภาษาไทยบางคำสามารถสื่อความหมายได้ชัดเจนโดยไม่จำเป็นต้องหารากศัพท์

3. การพัฒนาระบบจำแนกข้อความอัตโนมัติ ควรพัฒนาให้ครอบคลุมและสามารถใช้งานได้กับลักษณะข้อความที่หลากหลายรูปแบบ

กิตติกรรมประกาศ

งานวิจัยนี้ได้รับทุนอุดหนุนการวิจัยงบประมาณเงินรายได้ ประจำปีงบประมาณ 2557 และทุนอุดหนุนสนับสนุนการวิจัย งบประมาณแผ่นดิน ประจำปีงบประมาณ 2556 มหาวิทยาลัยมหาสารคาม

เอกสารอ้างอิง

1. วนิศา สถาพรวจนา. การทำเหมืองข้อความและออนโทโลจีในการจำแนกความสนใจของนักท่องเที่ยวต่อการท่องเที่ยวในประเทศไทย. สารนิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต. กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ; 2553.
2. Fan W, Wallace L, Rich S. Tapping The Power of Text Mining. Communication of the ACM 2006; 49: 76-82.
3. Thomas W. Data and Text Mining: A business Application Approach. 1st. United States: Prentice Hall; 2004.
4. H. Cherfi, A. Napoli, Y. Toussaint. Towards a text mining methodology using association rule extraction. Soft Computing 2006; 10: 431-441.
5. นิเวศ จิระวิชิตชัย. แบบจำลองการจำแนกเอกสารภาษาไทยอัตโนมัติ. วารสารวิชาการเทคโนโลยีอุตสาหกรรม 2556; 9(1): 141-149.
6. Aas K, Eikvil L. Text Categorization: a Survey. 1st. Norway: Norwegian Computing Center; 1999.
7. ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ. LexTo: เล็กซ์โตโปรแกรมตัดคำสำหรับข้อความภาษาไทย. [ออนไลน์]. 2546 [สืบค้นเมื่อ 10 มกราคม 2557]; ได้จาก http://www.sansarn.com/lexto.
8. Jaruskulchai C. An Automatic Indexing for Thai Text Retrieval. (Ph.D. thesis) USA: Grorge Washington University; 1998.
9. นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์, พยุง มีสัจ. การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ. NIDA Development Journal 2553; 51(3): 187-205.

10. ณิชาพร สุระ. การจำแนกหมวดหมู่เอกสารภาษาไทยอัตโนมัติโดยใช้อัลกอริทึม FPTC วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต. กรุงเทพฯ: สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ; 2549.
11. ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ. พจนานุกรมอิเล็กทรอนนิกส์ไทย-อังกฤษ เล็กซิตรอน. [ออนไลน์]. 2537 [สืบค้นเมื่อ 20 ธันวาคม 2556]; ได้จาก http://lexitron.nectec.or.th
12. Incham V. Automatic Thai Document Categorization System using SVM with Language Processing. (Ph.D. thesis) Bangkok: Kasetsart University; 2005.
13. Chirawichitchai N, Sa-nguansat P, Meesad P. A Comparative Study on Term Weight Techniques for Thai Document Categorization. Journal of Science Ladkrabang 2010; 19(1): 26-40.
14. Chirawichitchai N, Sa-nguansat P, Meesad P. An Experimental Study on Feature Reduction Techniques and Classification Algorithms of Thai Documents. Journal of Science Ladkrabang 2009; 18(2): 11-24.
15. Haruechaiyasak C, Jitkrittum W, Sangkeettrakarn C, Damrongrat C. Implementing News Article Category Browsing Based on Text Categorization Technique. International Conference on Web Intelligence and Intelligent Agent Technology; 9 - 12 December 2008; Sydney, New South Wales, Australia. WIIAT; 2008. pp. 143-146.
16. Yin S, Wang G, Qiu Y, Zhang W. Research and Implement of Classification Algorithm on Web Text Mining. Third International Conference on Semantics, Knowledge and Grid; 29 - 31 October 2007; Shan Xi, Chaina. SKG; 2007. pp. 446-449.
17. Huang H, Liu Q, Wu L, Huang T, Yuan S. The Application Research of Topic Word List in Text Automatic Classification. Second International Symposium on Knowledge Acquisition and Modeling; 30 November - 1 December 2009; Wuhan, Chaina. KAM; 2009. pp. 111-114.
18. Deng J, Hu J, Chi H, Wu J. An Improved Fuzzy Clustering Method for Text Mining. Second International Conference on Networks Security, Wireless Communications and Trusted Computing; 24 – 25 April 2010; Wuhan, Hubei, Chaina. NSWCTC; 2010. pp. 65-69.