ขั้นตอนการลดเทอมคุณสมบัติซ้ำซ้อน สำหรับการจำแนกบทเพลงไทย

The Algorithm of Feature Redundancy Reduction for Thai Lyrics Classification

ณัฐภัทร แก้วรัตนภัทร์[1] ***Nutthapat Kaewrattanapat***

บทคัดย่อ

จากอดีตจนถึงปัจจุบันมนุษย์ได้มีการใช้บทเพลงในการถ่ายทอดความรู้สึกนึกคิดและอารมณ์ต่าง ๆ ซึ่งได้รับความนิยมอย่างมากในการใช้สื่อสารลักษณะหนึ่งของมนุษย์ โดยมีการเลือกใช้คำ ประโยค หรือวลี ที่มีความสอดคล้องกับอารมณ์ที่จะถ่ายทอดให้แก่ผู้รับสาร จากการสำรวจพบว่าในเว็บไซต์และสื่อสังคมออนไลน์ต่าง ๆ ได้มีการเผยแพร่และแลกเปลี่ยนข้อมูลบทเพลงจำนวนมหาศาล และในการค้นคืนข้อมูลบทเพลงนั้น ยังไม่สามารถค้นคืนในลักษณะเชิงอารมณ์ (Sentiment) ของบทเพลงได้ เช่น หากต้องการค้นคืนบทเพลงที่มีอารมณ์รัก (Love Song) เมื่อใช้คำสำคัญ “บทเพลง+รัก” เครื่องมือจักรกลค้นคืน (Search Engine) จะค้นข้อมูลที่มีคำว่า “บทเพลง” และคำว่า “รัก” โดยแท้จริงแล้วจุดมุ่งหมายของผู้ค้นคืน คือ บทเพลงที่มีอารมณ์รัก ทำให้ผู้ค้นได้รับข้อมูลสนองกลับที่ไม่ตรงจุดประสงค์นั่นเองนอกจากนี้ในการจำแนกประเภทบทเพลงมีการเกิดคุณลักษณะที่ซ้ำซ้อนเนื่องจากบทเพลงประเภทบทเพลงรัก และบทเพลงเศร้ามีการใช้คำที่คล้ายคลึงกันจำนวนมาก ทำให้เกิดความถูกต้องในการจำแนกต่ำ

[1] อาจารย์ ประจำสาขาวิชาสารสนเทศศาสตร์ แขนงวิชาระบบสารสนเทศเพื่อการจัดการ ภาควิชาสังคมศาสตร์ คณะมนุษยศาสตร์และสังคมศาสตร์ มหาวิทยาลัยราชภัฏสวนสุนันทา; Lecturer, Information Science Program (Management Information System), Department of Social Sciences, Faculty of Humanities and Social Sciences, Suan Sunandha Rajabhat University

รับต้นฉบับ 5 กรกฎาคม 2559 ปรับแก้ไขตามข้อเสนอแนะของผู้ทรงคุณวุฒิ 12 กันยายน 2559
รับลงตีพิมพ์ 19 กันยายน 2559

การวิจัยครั้งนี้ได้ทำการศึกษาและนำเสนอขั้นตอนการลดเทอมคุณสมบัติซ้ำซ้อนสำหรับการจำแนกบทเพลงไทย โดยใช้วิธีการลดความซ้ำซ้อนด้วยการลำดับค่า TF*IDF และพยากรณ์ด้วยการวัดความคล้ายคลึงโคไซน์ โดยจากการทดลองทำให้ทราบว่าการลดคุณสมบัติ (Feature Reduction) ของบทเพลงทำให้การพยากรณ์ประเภทของบทเพลงมีความถูกต้องสูงที่สุด คือ การลดคุณสมบัติลงจำนวนร้อยละ 70 มีความถูกต้องร้อยละ 75 โดยใช้เวลาในการประมวลผล คือ 0.76 วินาที ซึ่งลดระยะเวลาในการประมวลผลลดลง 1.54 วินาที ทำให้เกิดประสิทธิภาพทั้งด้านความถูกต้องและด้านเวลาในการประมวลผล

Abstract

From the past to the present time, humans have used songs for conveying their emotions, thoughts, and sentiments. People used this popular communication approach by selecting words, sentences or phrases related to their emotions and transmitting them to receivers. It was discovered that websites and social media have presented and exchanged a large number of songs and lyrics, but the song retrieving through their sentimentality were impossible. For instance, if you wanted to retrieve love songs and you used key words “lyrics + love,” the computer search engines would render to you the entries with the words “songs + love.” Actually, the searchers were expecting love emotion oriented songs. This resulted in the failure to respond to the searcher needs. In addition, the lyrics classification produced redundancy as love lyrics and lamenting lyrics used a large number of similar words. This led to low accuracy of song classification.

This study proposed and presented the algorithm of feature redundancy reduction for Thai lyrics classification, using TF*IDF feature reduction and predicting the Cosine Similarity. Through the experiment, the results indicated that the feature reduction achieved the highest accuracy of lyrics classification. Seventy percents of the feature reduction produced 75% accuracy. Its processing time is 0.76 seconds, and the duration of processing time was decreased to 1.54 seconds. In

addition, the feature reduction reduced processing time and increased the classification accuracy.

คำสำคัญ: การลดคุณสมบัติ การวิเคราะห์เชิงอารมณ์ การจำแนกเอกสารภาษาไทย ภาษาศาสตร์เชิงคำนวณ

Keywords: Feature Reduction, Sentiment Analysis, Thai Document Classification, Computational Linguistic

บทนำ

จากอดีตจนถึงปัจจุบัน มนุษย์ได้มีการใช้บทเพลงในการถ่ายทอดความรู้สึกนึกคิดและอารมณ์ต่าง ๆ ซึ่งได้รับความนิยมอย่างมากในการใช้สื่อสารลักษณะหนึ่งของมนุษย์ โดยมีการเลือกใช้คำ ประโยค หรือวลี ที่มีความสอดคล้องกับอารมณ์ที่จะถ่ายทอดให้แก่ผู้รับสารได้ จากการสำรวจพบว่า ในเว็บไซต์และสื่อสังคมออนไลน์ต่าง ๆ ได้มีการเผยแพร่และแลกเปลี่ยนข้อมูลบทเพลงจำนวนมหาศาล และในการค้นคืนข้อมูลบทเพลงนั้น ยังไม่สามารถค้นคืนในลักษณะเชิงอารมณ์ (Sentiment) ของบทเพลงได้ เช่น หากต้องการค้นคืนบทเพลงที่มีอารมณ์รัก (Love Song) เมื่อใช้คำสำคัญ “บทเพลง+รัก” เครื่องมือจักรกลค้นคืน (Search engine) จะค้นข้อมูลที่มีคำว่า “บทเพลง” และคำว่า “รัก” โดยแท้จริงแล้วจุดมุ่งหมายของผู้ค้นคืน คือ บทเพลงที่มีอารมณ์รัก ทำให้ผู้ค้นได้รับข้อมูลสนองกลับที่ไม่ตรงจุดประสงค์นั่นเอง

จากปัญหาดังกล่าว จึงเกิดแนวคิดในการพัฒนาระบบการจำแนกบทเพลงขึ้นเพื่อแก้ไขปัญหาการค้นคืนบทเพลงเชิงอารมณ์ โดยระบบที่พัฒนาขึ้นจะใช้ทดสอบกับบทเพลงที่มีเนื้อเพลงเป็นภาษาไทย จำนวน 2 อารมณ์ คือ อารมณ์รัก และ อารมณ์เศร้า เนื่องจากภาษาไทยเป็นภาษาประจำชาติที่คนไทยนิยมใช้กันแพร่หลายมากกว่าภาษาอื่น ทำให้มีบทเพลงไทยจำนวนมากตั้งแต่อดีตจนถึงปัจจุบัน และภาษาไทยมีความยากในการตัดคำเนื่องจากภาษาไทยไม่มีการเว้นวรรคระหว่างคำ ซึ่งต่างจากภาษาอังกฤษ และภาษาไทยยังเป็นภาษาที่มีคำฟุ่มเฟือยจำนวนมากทำให้ยากต่อการเปรียบเทียบรูปแบบ (Kunnu, & Kaewrattanapat 2013) นอกจากนี้ การจำแนกประเภทบทเพลงมีการเกิดคุณลักษณะที่ซ้ำซ้อนเนื่องจากบทเพลงประเภทบทเพลงรัก และบทเพลงเศร้ามีการใช้คำที่คล้ายคลึงกัน

จำนวนมาก จึงทำให้การจำแนกบทเพลงไทยตามอารมณ์ความรู้สึกมีประสิทธิภาพต่ำ ซึ่งงานวิจัยที่ได้ดำเนินการนี้เป็นการเสนอขั้นตอนการลดคุณสมบัติ (Feature reduction) ของคำจากเนื้อเพลงไทยที่มีความซ้ำซ้อนกันในแต่ละกลุ่มผสมผสานกับวิธีการวัดความคล้ายคลึงเชิงมุม (Cosine similarity) ซึ่งจะทำให้ลดจำนวนคุณลักษณะของคำลงจำนวนมากส่งผลให้เกิดประสิทธิภาพทางด้านความเร็วในการจำแนกและลดหน่วยความจำในการจัดเก็บคลังคำเนื้อเพลงไทยได้

ดังนั้น การพัฒนาขั้นตอนการลดคุณสมบัติซ้ำซ้อนสำหรับการจำแนกบทเพลงไทยนั้น จะช่วยให้การจำแนกกลุ่มบทเพลงตามอารมณ์มีความรวดเร็วมากยิ่งขึ้น และเป็นการเพิ่มมุมมองในการค้นคืนเอกสารเชิงอารมณ์โดยระบบดังกล่าวสามารถเพิ่มพูนการเรียนรู้รูปแบบของคำในแต่ละอารมณ์เพลงของตนเองได้เรื่อยๆ (Self-learning) นอกจากนี้สามารถใช้ประยุกต์ในการวัดความคล้ายคลึงในระบบต่างๆ เช่น การแบ่งกลุ่มเอกสารและการค้นหาเอกสารที่เกี่ยวข้องกับที่ผู้ใช้กำหนด

วัตถุประสงค์

1. เพื่อพัฒนาขั้นตอนการลดคุณสมบัติซ้ำซ้อนสำหรับการจำแนกบทเพลงไทย
2. เพื่อประเมินประสิทธิภาพของขั้นตอนการลดคุณสมบัติซ้ำซ้อนสำหรับการจำแนกบทเพลงไทย

ผลที่คาดว่าจะได้รับ

1. สามารถจำแนกอารมณ์ของบทเพลงได้อย่างมีประสิทธิภาพ และลดจำนวนของคำที่ใช้ในการเปรียบเทียบ
2. เพิ่มความสามารถแก่ระบบจักรกลค้นคืนให้สามารถค้นคืนข้อมูลเชิงอารมณ์ได้

การทบทวนวรรณกรรม

จากการศึกษาและพัฒนาขั้นตอนการลดคุณสมบัติสำหรับการจำแนกบทเพลงไทย ผู้วิจัยได้ค้นคว้าเอกสารและงานวิจัยที่เกี่ยวข้องเพื่อให้สามารถจัดทำระบบการจำแนกบทเพลงได้อย่างมีประสิทธิภาพ โดยมีเอกสารที่เกี่ยวข้องดังต่อไปนี้

1. การตัดคำ (Words segmentation)

การตัดคำภาษาไทย (Thai words segmentation) ได้รับการพัฒนาขึ้นมาโดยใช้วิธีการต่างๆ ที่ต่างกัน เนื่องจากการตัดคำเป็นกระบวนการพื้นฐานของการประมวลผลภาษาธรรมชาติ เช่น การวิเคราะห์เสียงพูด การตัดคำภาษาไทยเองก็เช่นกัน ได้มีผู้คิดค้นวิธีที่จะแยกคำแต่ละคำออกจากประโยคซึ่งมีการเขียนติดกันไปอย่างต่อเนื่องทั้งประโยค ในงานวิจัยนี้จะกล่าวถึงการตัดคำโดยอาศัยอักขรวิธี เป็นหลักการพื้นฐานการประสมคำ

1) การใช้กฎ (Rule Based segmentation)

การตัดคำโดยการตรวจสอบกฎเกณฑ์ทางอักขรวิธีที่กำหนดลักษณะการประสมอักษรลักษณะการเว้นวรรค และการขึ้นย่อหน้า เพื่อใช้เป็นเกณฑ์ในการกำหนดขอบเขตของคำ วิธีการนี้จะมีข้อจำกัดในการทำงาน คือ ความถูกต้องของการตัดคำในระดับพยางค์สูงแต่ความถูกต้องของการตัดคำค่อนข้างต่ำ แต่ข้อดีคือมีความรวดเร็วในการทำงานและใช้ทรัพยากรน้อย

2) การใช้พจนานุกรม (Dictionary approach)

การตัดคำโดยพจนานุกรมเป็นการตัดคำโดยใช้สายอักขระ (String) มาเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรม ซึ่งวิธีนี้จะต้องทำการจัดเก็บคำไว้ในพจนานุกรม วิธีนี้ทำให้ได้ความถูกต้องในการตัดคำสูงกว่าการใช้กฎแต่จะใช้เวลาในการทำงานมากกว่า

3) การใช้คลังข้อความ (Corpus-based approach)

การตัดคำโดยใช้คลังข้อมูลเป็นการตัดคำโดยนำวิธีการทางสถิติเข้ามาใช้ในการประมวลภาษา โดยใช้คลังข้อมูลทางภาษาเป็นฐานความรู้เก็บค่าความถี่ที่ใช้ในการตัดคำ ซึ่งการตัดคำโดยใช้คลังข้อมูลแบ่งออกเป็น 2 วิธี คือ การตัดคำโดยอาศัยความน่าจะเป็น และวิธีการตัดคำโดยอาศัยคุณลักษณะของคำ

4) เทคนิควิธีการเทียบคำที่ยาวที่สุด (Longest word pattern matching)

วิธีนี้จะทำการตรวจสอบสายอักขระ (String) ที่นำเข้ามาจากซ้ายไปขวา จากนั้นนำไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรม หากตรวจสอบพบว่าพบพยางค์มากกว่า 1 พยางค์ในพจนานุกรม จะทำการเลือกพยางค์ที่ยาวที่สุดแล้วทำต่อไปเรื่อยๆ จนจบสายอักขระ

ตัวอย่างคำว่า “ดวงดาว” การตัดคำโดยวิธีนี้จะนำสายอักขระไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรมจะพบคำว่า ด, ดว และคำว่า ดวง ส่วนคำว่า ดวงด ไม่พบอยู่ในพจนานุกรม ดังนั้นจึงได้คำว่า ดวง ซึ่งเป็นคำที่ยาวที่สุดที่หาพบ ส่วนที่เหลือคือ ดาว เมื่อนำไปค้นในพจนานุกรมจะได้ว่า ด, ดา, ดาว ดังนั้น จึงเลือกคำว่า ดาว คำที่ได้จากการตัดคำโดยวิธีการนี้จึงเป็น “ดวงดาว” วิธีการนี้ให้ความถูกต้องหลังการตัดคำสูงกว่าวิธีการอื่นโดยเฉพาะใช้ร่วมกับวิธีย้อนรอยกลับ

2. การกำจัดคำฟุ่มเฟือย (Stop-word removal)

เป็นการนำคำที่ไม่มีนัยสำคัญออกโดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลง คำที่ไม่มีนัยสำคัญในที่นี้หมายถึงคำที่ใช้กันโดยทั่วไปที่ไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง ตัวอย่างเช่น คำบุพบทเป็นคำที่ใช้เชื่อมคำหรือกลุ่มคำให้สัมพันธ์กัน คำสันธานเป็นคำที่ทำหน้าที่เชื่อมประโยคกับประโยค คำสรรพนามเป็นคำที่ใช้แทนคำนามที่กล่าวถึงมาแล้วในประโยค เป็นต้น คำหยุดมักเป็นคำที่ปรากฏขึ้นบ่อยครั้งในเอกสารและปรากฏในเอกสารเกือบทุกฉบับ จึงถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการค้นคืนหรือการจำแนกหมวดหมู่ ดังนั้น การกำจัดคำหยุดจึงเป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี เพื่อกำจัดคุณลักษณะที่ไม่เป็นประโยชน์และลดขนาดของดัชนีลง ซึ่งจะช่วยประหยัดทั้งพื้นที่และเวลาในการประมวลผล ตัวอย่างคำหยุด เช่น “ที่” “ว่า” “และ” “จะ” “ได้” “ของ” และ “ให้” เป็นต้น

3. แบบจำลองเวกเตอร์สเปซของเนื้อเพลงไทย

เป็นการจัดรูปแบบของเอกสารหรือตัวแทนของเอกสาร (Document representation) เพื่อให้ง่ายต่อการประมวลผล ซึ่งรูปแบบจะมีลักษณะของการแสดงความสัมพันธ์ระหว่างคำในเอกสารทั้งหมด ซึ่งคำที่ได้นั้นต้องผ่านกระบวนการทำให้เป็นบรรทัดฐานแล้ว เช่น การตัดคำฟุ่มเฟือยออกไป รวมถึงการให้น้ำหนักกับคำเหล่านั้น บางครั้งเรียกตามรูปแบบของเอกสารที่ปรากฏว่าถุงของคำ (Bag of words) (Hiemstra, 2000)

เทคนิคการกำหนดค่าความสำคัญให้กับคำในการแทนเอกสารเป็นแบบจำลองเวคเตอร์ (Vector-Space model) เช่น

1. ค่าความถี่ของเทอมคุณลักษณะ (Term frequency: TF) คือ จำนวนครั้งที่เทอมคุณลักษณะเกิดขึ้นในทุกๆ เอกสาร ดังตารางที่ 1

ตารางที่ 1 แสดงตัวอย่างค่าความถี่ของเทอมคุณลักษณะ (TF) ของเอกสาร

เทอมคุณลักษณะที่ปรากฏ	จำนวนความถี่ของเทอมคุณลักษณะที่ปรากฏคำภายในเอกสารแต่ละฉบับ (หน่วย: คำ) TF	
	เอกสารฉบับที่ 1	เอกสารฉบับที่ 2
รัก	300	400
ไม่รัก	150	200
ท้อ	30	110
คิดถึง	170	190

2. ค่าบรรทัดฐานความถี่ของเทอมคุณลักษณะ (Normalized term frequency) คำนวณตามสูตร ดังนี้

$$Nomalized\ TF = \frac{TF_i}{\sum TF}$$

โดยที่ TF_i : ค่า TF ตัวที่ i ของเทอมคุณลักษณะที่ปรากฏในเอกสาร, $\sum TF$: ค่าผลรวมของค่า TF_i ในเอกสาร, $Nomalized\ TF$: ค่าบรรทัดฐานความถี่ของเทอมคุณลักษณะ

ตารางที่ 2 แสดงตัวอย่างค่าบรรทัดฐานความถี่ของเทอมคุณลักษณะ

เทอมคุณลักษณะที่ปรากฏ	จำนวนความถี่ของเทอมคุณลักษณะ			
	เอกสารฉบับที่ 1		เอกสารฉบับที่ 2	
	TF	Normalized TF	TF	Normalized TF
รัก	300	0.397	240	0.320
ไม่รัก	150	0.199	200	0.270
ท้อ	30	0.040	110	0.150
คิดถึง	270	0.357	190	0.260
พิศวาส	5	0.007	0	0
รวม	755	1	740	1

3) ค่าความถี่ผกผันของเอกสาร (Inverse document frequency: IDF) โดยที่ค่า IDF ทำให้ทราบว่าเทอมคุณลักษณะมีการปรากฏในเอกสารอย่างทั่วถึงหรือไม่ เมื่อเทอมคุณลักษณะปรากฏในเอกสารทุกฉบับ ค่า IDF จะมีค่าเท่ากับ 1 และหากเทอมคุณลักษณะปรากฏในเอกสารบางฉบับ ค่า IDF จะมีค่ามากขึ้นห่างจากค่า 1 เรื่อยๆ ในทิศทางบวก ตามความเบาบางที่ปรากฏในเอกสาร (Aizawa, 2000) โดยค่าความถี่ผกผันของเอกสารคำนวณตามสูตรดังนี้

$$IDF(w) = log_2 \frac{N}{n_p}$$

โดยที่ w : เทอมคุณลักษณะหรือคำ, N : จำนวนเอกสารทั้งหมดในฐานข้อมูล, n_p : จำนวนเอกสารในฐานข้อมูลที่พบคำ w

ตารางที่ 3 แสดงตัวอย่างค่าความถี่ผกผันของเอกสาร

เทอมคุณลักษณะที่ปรากฏ	ค่าความถี่ผกผันของเอกสาร		
	ปรากฏในเอกสารฉบับที่ 1	ปรากฏในเอกสารฉบับที่ 2	IDF
รัก	Y	Y	1
ไม่รัก	Y	Y	1
ท้อ	Y	Y	1
คิดถึง	Y	Y	1
พิศวาส	Y	N	1.301

Y: ปรากฏและ N: ไม่ปรากฏ

ค่าผลคูณของค่าบรรทัดฐานความถี่ของเทอมคุณลักษณะกับค่าความถี่ผกผันของเอกสาร (TF*IDF) เป็นเทคนิคการกำหนดค่าความสำคัญให้กับคำในการแทนเอกสารเป็นแบบจำลองเวคเตอร์ (Vector-space model) คือ ค่าผลคูณระหว่าง TF (Term frequency) และ IDF (Inverse Document frequency) ซึ่งเป็นการเปรียบเทียบความถี่ของคำในเอกสารกับความถี่ของคำในเอกสารอื่น การปรากฏคำนั้นในส่วนต่างๆ ของเอกสาร (Aizawa, 2001) คำนวณตามสูตร ดังนี้

$$TF * IDF = F(w,D) * log_2 \frac{N}{n_p}$$

โดยที่ D : เอกสาร, w : คำ, F(w,D) : ความถี่ที่พบคำ w ในเอกสาร D, n_p : จำนวนเอกสารในฐานข้อมูลที่พบคำ w, N : จำนวนเอกสารทั้งหมดในฐานข้อมูล

ตารางที่ 4 แสดงตัวอย่างค่าผลคูณของค่าบรรทัดฐานความถี่ของเทอมคุณลักษณะกับค่าความถี่ผกผันของเอกสาร (TF*IDF)

เทอมคุณลักษณะที่ปรากฏ	จำนวนความถี่ของเทอมคุณลักษณะ							
	เอกสารฉบับที่ 1				เอกสารฉบับที่ 2			
	TF	Normalized TF	IDF	TF*IDF	TF	Normalized TF	IDF	TF*IDF
รัก	300	0.397	1	0.397	240	0.320	1	0.320
ไม่รัก	150	0.199	1	0.199	200	0.270	1	0.270
ท้อ	30	0.040	1	0.040	110	0.150	1	0.150
คิดถึง	270	0.357	1	0.357	190	0.260	1	0.260
พิศวาส	5	0.007	1.301	0.009	0	0	1.301	0

4. การวัดความคล้ายคลึงเชิงมุม (Cosine similarity method)

การวัดความคล้ายคลึงของเอกสารด้วยวิธีการวัดความคล้ายคลึงเชิงมุมนั้นเป็นวิธีการเปรียบเทียบความคล้ายคลึงของเอกสารสองเอกสาร โดยแต่ละเอกสารจะถูกแทนด้วยเวกเตอร์ขนาดเอ็น (N-Dimensional Vector) ซึ่งเก็บค่าน้ำหนักแต่ละค่าในเอกสารนั้น การเปรียบเทียบความคล้ายคลึงของเอกสารจะเปรียบเทียบโดยดูจากมุมโคไซน์ของมุมระหว่าง 2 เวกเตอร์ของเอกสาร หากเอกสารทั้งสองเอกสารคล้ายคลึงกันมาก เวกเตอร์ของเอกสารทั้ง 2 จะทับกันเกือบสนิท มุมจึงมีค่าน้อย ค่าโคไซน์ที่ได้จะมีค่ามาก (Kunnu, & Kaewrattanapat 2013)

วิธีการดำเนินการวิจัย

ในการทดลองขั้นตอนการลดเทอมคุณสมบัติซ้ำซ้อนสำหรับการจำแนกบทเพลงไทยนั้น มีกระบวนการดังต่อไปนี้

1. การเตรียมข้อมูล

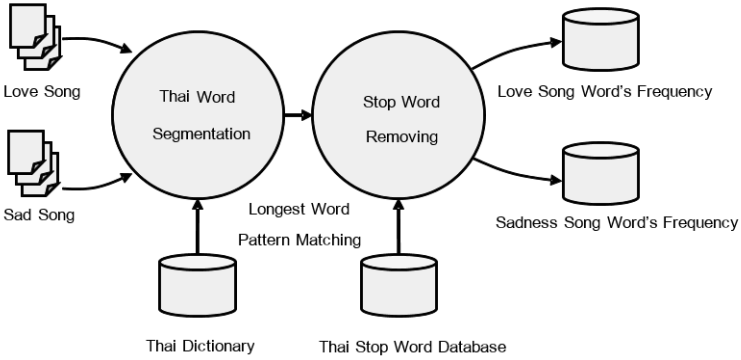
1) ขอบเขตด้านข้อมูลการฝึกสอน (Training set) ข้อมูลที่นำมาใช้ในการเรียนรู้เป็นบทเพลงไทยในรูปแบบไฟล์อิเล็กทรอนิกส์ (Plain text) ซึ่งนำมาจำแนกออกเป็นกลุ่ม ได้แก่ บทเพลงรักและบทเพลงเศร้า จำนวน 2,000 บทเพลงในแต่ละประเภท และใช้นักภาษาศาสตร์ในการจำแนกอารมณ์ของบทเพลง

2) ขอบเขตด้านการทดสอบ (Test set) ข้อมูลที่นำมาใช้ในการทดสอบเป็นบทเพลงไทยในรูปแบบไฟล์อิเล็กทรอนิกส์ ได้แก่ บทเพลงรักและบทเพลงเศร้า จำนวน 200 บทเพลงในแต่ละประเภท และเลือกใช้วิธีการวัดค่าความถูกต้อง (Accuracy) เพื่อวัดประสิทธิภาพสำหรับการประเมินผลการจำแนกกลุ่มข้อมูล โดยใช้ตาราง Confusion matrix แสดงค่าข้อมูลของตัวชี้วัดเหล่านี้ขึ้นอยู่กับจำนวนของข้อมูลที่สามารถพยากรณ์ถูกต้อง และพยากรณ์ไม่ถูกต้องด้วย ซึ่งวิธีการนี้จะทำการวัดได้ทั้งค่าความถูกต้อง และค่าความผิดพลาดของการจำแนกกลุ่ม

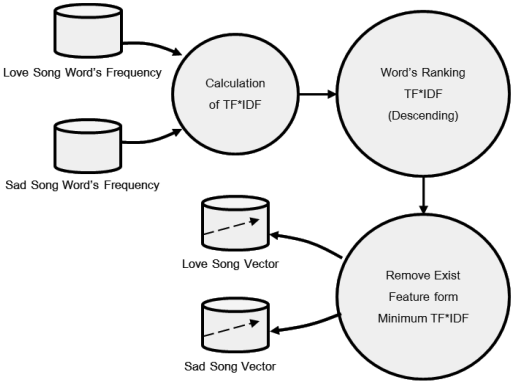
3) ฐานข้อมูลคำศัพท์ที่มีลักษณะเป็นคำหยุด (Stop word database) ฐานข้อมูลคำศัพท์ที่มีลักษณะเป็นคำหยุดประกอบไปด้วยคำหยุดในภาษาไทย จำนวน 415 คำ ซึ่งคลังคำหยุดจะนำมาใช้กำจัดคำฟุ่มเฟือยที่เกิดขึ้นบ่อยครั้ง โดยหากไม่มีการกำจัดคำเหล่านี้ จะทำให้เกิดการโน้มเอียงของผลการพยากรณ์ผิดพลาดได้สูง

2. การประมวลผลเบื้องต้น (Data preprocessing)

ในการประมวลผลเบื้องต้นจะมีการตัดคำไทยด้วยพจนานุกรมไทยโดยใช้เทคนิควิธีการเทียบคำที่ยาวที่สุด (Longest word pattern matching) จากนั้นจะเข้าสู่ขั้นตอนการกำจัดคำหยุด (Stop word) ด้วยฐานข้อมูลคำหยุด และนับความถี่ของคำศัพท์ที่คงเหลืออยู่จากบทเพลง ดังรูปที่ 1



รูปที่ 1 แสดงกระบวนการประมวลผลเบื้องต้น (Data Preprocessing)



รูปที่ 2 แสดงกระบวนการกำจัดคำซ้ำด้วยค่า TF*IDF

จากนั้นจะเข้าสู่กระบวนการกำจัดคำซ้ำโดยใช้การคำนวนค่า TF*IDF และจัดเรียงค่าที่ได้จากค่ามากไปหาค่าน้อย (Descending) โดยจะกำจัดคำซ้ำที่มีค่า TF*IDF ต่ำ

3. การพยากรณ์การจำแนก (Data Classification)

การเปรียบเทียบความคล้ายคลึงของเอกสารจะเปรียบเทียบโดยดูจากมุมโคไซน์ของมุมระหว่าง 2 เวกเตอร์ของเอกสาร หากเอกสารทั้งสองเอกสารคล้ายคลึงกันมาก เวกเตอร์ของเอกสารทั้ง 2 จะทับกันเกือบสนิท มุมจึงมีค่าน้อย ค่าโคไซน์ที่ได้จะมีค่ามาก (Haifeng, Zheng, Ahmad & HuiTian, 2014) โดยการวัดค่าความเหมือนของเอกสารซึ่งพิจารณาค่าโคไซน์จากสมการ

$$Similarity = COS\theta = \frac{D \cdot \bar{D}}{||D||||\bar{D}||}$$

$$D \cdot \bar{D} = \sum_{i=1}^{n} (d_i * \bar{d}_i)$$

$$||D|| = \sqrt{\sum_{i}^{n} d_i^2}$$

$$||\bar{D}|| = \sqrt{\sum_{i}^{n} \bar{d}_i^2}$$

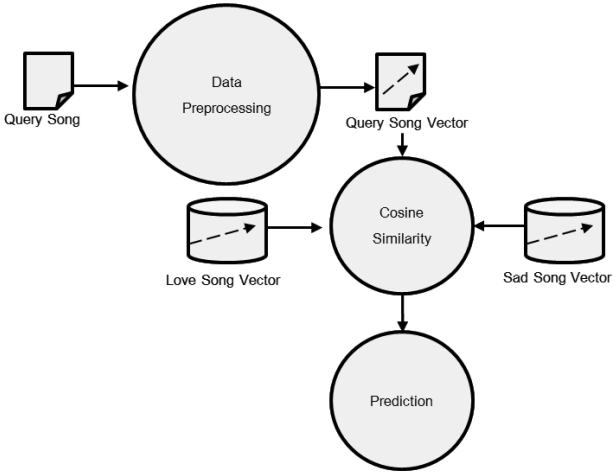
โดยที่ D คือ เอกสารที่ต้องการวัดค่าความคล้ายคลึง

$\bar{D}$ คือ เอกสารเทียบเคียงความคล้ายคลึง

n คือ จำนวนเทอมของคำที่ปรากฏในเอกสาร D และ $\bar{D}$

d_i คือ ค่า TF*IDF ของเทอมตัวที่ i ที่ปรากฏในเอกสาร D

$\bar{d}_i$ คือ ค่า TF*IDF ของเทอมตัวที่ i ที่ปรากฏในเอกสาร $\bar{D}$



รูปที่ 3 แสดงกระบวนการพยากรณ์การจำแนก (Data classification)

4. การประเมินผล (Data evaluation)

ในการทดสอบครั้งนี้ เลือกใช้วิธีการวัดค่าความถูกต้อง (Accuracy) เพื่อวัดประสิทธิภาพมาตรฐานสำหรับการประเมินผลการจำแนกกลุ่มข้อมูลที่ใช้ในการทดลองเปรียบเทียบ โดยค่าข้อมูลของตัวชี้วัดเหล่านี้ขึ้นอยู่กับจำนวนของข้อมูลที่สามารถพยากรณ์ถูกต้อง และพยากรณ์ไม่ถูกต้องด้วยตาราง Confusion matrix ซึ่งวิธีการนี้จะทำการวัดได้ทั้งค่าความถูกต้อง และค่าความผิดพลาดของการจำแนกกลุ่ม

ตารางที่ 5 การวัดประสิทธิภาพด้วย Confusion matrix

CLASS True		Predicted Class	
		False	
Actual Class	True	TP	FN
	False	FP	TN

โดยสามารถคำนวณได้จากสมการ

$$\%Accuracy = \left(\frac{TP+TN}{TP+TN+FN+FP}\right) * 100$$

โดยที่ TP (True positives) คือ จำนวนของการทดลองที่ทำนายถูกต้องกรณีที่คลาสจริงเป็นจริง

TN (True negatives) คือ จำนวนของการทดลองที่ทำนายถูกต้องกรณีที่คลาสจริงเป็นเท็จ

FN (False negatives) คือ จำนวนของการทดลองที่ทำนายผิดกรณีที่คลาสจริงเป็นจริง

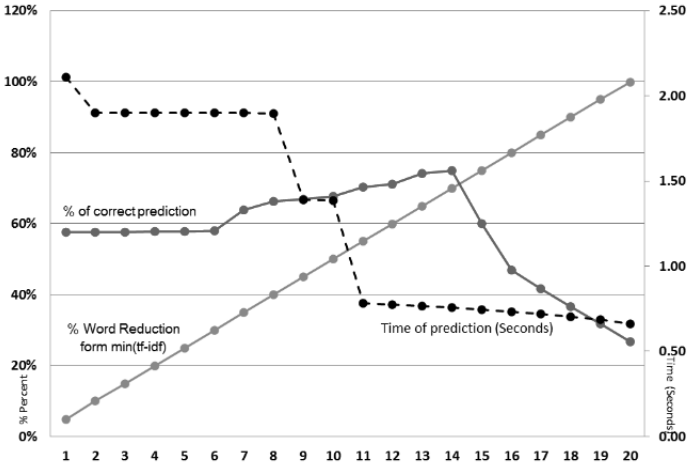
FP (False positives) คือ จำนวนของการทดลองที่ทำนายผิดกรณีที่คลาสจริงเป็นเท็จ

นอกจากนี้ผู้วิจัยได้ใช้วิธีการประเมินประสิทธิภาพความแม่นยำจากการพิสูจน์แบบไขว้ 10 รอบ (10 Fold cross validation) ในการเปรียบเทียบผลการทดลองเพื่อยืนยัน

ประสิทธิภาพความแม่นยำของผลการทดลองด้วย ซึ่งการพิสูจน์แบบไขว้เป็นที่นิยมในการยืนยันผลการทดลองทางด้านการวัดประสิทธิภาพความถูกต้องของข้อมูล (Ron Kohavi.,1995)

ผลการวิจัย

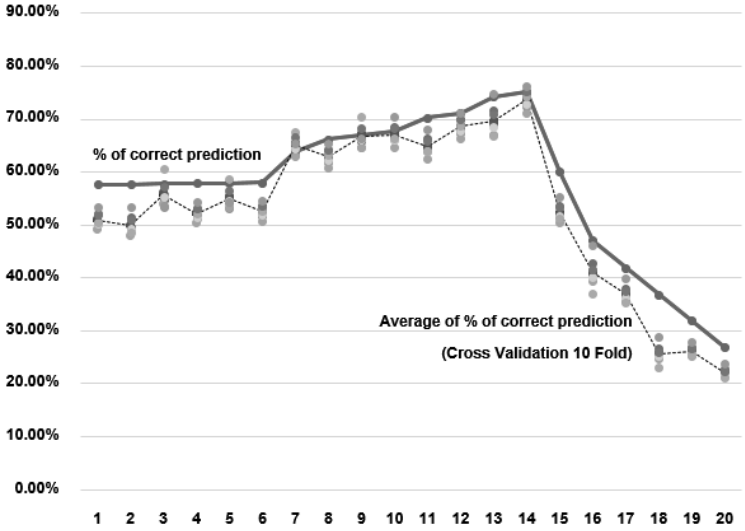
การวิจัยครั้งนี้ได้ทำการศึกษาและนำเสนอขั้นตอนการลดเทอมคุณสมบัติซ้ำซ้อนสำหรับการจำแนกบทเพลงไทย โดยใช้วิธีการลดความซ้ำซ้อนด้วยการลำดับค่า TF*IDF และพยากรณ์ด้วยการวัดความคล้ายคลึงโคไซน์ซึ่งผลการทดลองจำนวน 20 ครั้ง ปรากฏใน รูปที่ 4 และ ตารางที่ 6



รูปที่ 4 แสดงการเปรียบเทียบผลการทดลองการลดความซ้ำซ้อนของคุณลักษณบทเพลงไทย

ตารางที่ 6 ผลการทดลองการลดความซ้ำซ้อนของคุณลักษณะบทเพลงไทย

ลำดับการทดลอง	ไม่มีการลดความซ้ำซ้อนของคุณลักษณะ			มีการลดความซ้ำซ้อนของคุณลักษณะ		
	% การลดความซ้ำซ้อน	ความถูกต้องในการจำแนก	เวลาในการพยากรณ์จำแนก (วินาที)	% การลดควาซ้ำซ้อน โดยพิจารณาจากค่า min(tf*idf)	ความถูกต้องในการจำแนก	เวลาในการพยากรณ์จำแนก (วินาที)
1	0	57%	2.30	5%	57.625%	2.11
2	0	57%	2.30	10%	57.626%	1.90
3	0	57%	2.30	15%	57.776%	1.90
4	0	57%	2.30	20%	57.976%	1.90
5	0	57%	2.30	25%	57.976%	1.90
6	0	57%	2.30	30%	58.093%	1.90
7	0	57%	2.30	35%	63.965%	1.90
8	0	57%	2.30	40%	66.271%	1.89
9	0	57%	2.30	45%	67.027%	1.39
10	0	57%	2.30	50%	67.677%	1.38
11	0	57%	2.30	55%	70.330%	0.78
12	0	57%	2.30	60%	71.111%	0.77
13	0	57%	2.30	65%	74.246%	0.76
14	0	57%	2.30	70%	75.092%	0.76
15	0	57%	2.30	75%	60.094%	0.74
16	0	57%	2.30	80%	47.097%	0.73
17	0	57%	2.30	85%	41.900%	0.72
18	0	57%	2.30	90%	36.903%	0.70
19	0	57%	2.30	95%	31.906%	0.68
20	0	57%	2.30	100%	26.909%	0.66



รูปที่ 5 แสดงการเปรียบเทียบผลการทดลองการระหว่างค่าความแม่นยำที่ได้จากข้อมูลทดสอบกับค่าเฉลี่ยความแม่นยำจากการพิสูจน์แบบไขว้ 10 รอบ

ตารางที่ 7 ผลการทดลองระหว่างค่าความแม่นยำที่ได้จากข้อมูลทดสอบกับค่าเฉลี่ยความแม่นยำจากการพิสูจน์แบบไขว้ 10 รอบ

% การลดความซ้ำซ้อน	% ความถูกต้อง	% ความถูกต้องในการจำแนก (Cross Validation K=10 Fold)										
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	ค่าเฉลี่ย
5%	57.63%	50.40%	50.22%	50.53%	52.10%	51.01%	49.10%	50.29%	50.63%	51.97%	53.31%	50.96%
10%	57.63%	49.49%	49.59%	49.66%	51.40%	49.98%	48.10%	49.27%	48.51%	50.91%	53.31%	50.02%
15%	57.78%	54.95%	53.52%	54.91%	55.71%	56.17%	54.06%	55.38%	60.73%	57.23%	53.73%	55.64%
20%	57.98%	51.63%	51.35%	51.75%	53.28%	52.29%	50.33%	51.56%	52.28%	53.28%	54.28%	52.20%
25%	57.98%	54.37%	53.26%	54.37%	55.40%	55.44%	53.35%	54.65%	58.68%	56.48%	54.28%	55.03%
30%	58.09%	51.94%	51.66%	52.06%	53.60%	52.61%	50.63%	51.87%	52.60%	53.60%	54.60%	52.52%
35%	63.97%	64.21%	63.31%	64.28%	65.78%	65.30%	62.85%	64.38%	67.52%	66.52%	65.52%	64.97%
40%	66.27%	62.32%	61.94%	62.46%	64.28%	63.15%	60.78%	62.26%	63.34%	64.34%	65.34%	63.02%
45%	67.03%	65.95%	64.78%	65.98%	67.34%	67.17%	64.65%	66.22%	70.43%	68.43%	66.43%	66.74%
50%	67.68%	66.36%	66.18%	66.54%	68.65%	67.14%	64.62%	66.19%	66.40%	68.40%	70.40%	67.09%
55%	70.33%	64.07%	63.91%	64.25%	66.29%	64.81%	62.38%	63.90%	64.03%	66.03%	68.03%	64.77%
60%	71.11%	67.85%	67.45%	68.00%	70.00%	68.74%	66.16%	67.77%	68.83%	70.03%	71.23%	68.61%
65%	74.25%	68.99%	69.28%	69.25%	71.79%	69.58%	66.97%	68.60%	66.89%	70.89%	74.89%	69.71%
70%	75.09%	72.95%	72.45%	73.10%	75.20%	73.94%	71.16%	72.90%	74.33%	75.33%	76.33%	73.77%
75%	60.09%	51.80%	51.77%	51.96%	53.68%	52.36%	50.39%	51.62%	51.34%	53.34%	55.34%	52.36%
80%	47.10%	40.72%	41.42%	40.95%	42.84%	40.83%	39.30%	40.26%	37.10%	41.60%	46.090%	41.11%
85%	41.90%	36.61%	36.81%	36.76%	38.14%	36.91%	35.52%	36.39%	35.30%	37.60%	39.90%	36.99%
90%	36.90%	25.50%	25.95%	25.65%	26.84%	25.56%	24.60%	25.20%	23.18%	26.04%	28.90%	25.74%
95%	31.91%	25.93%	25.97%	26.02%	26.92%	26.18%	25.20%	25.82%	25.45%	26.68%	27.90%	26.21%
100%	26.91%	21.92%	22.04%	22.01%	22.84%	22.09%	21.26%	21.78%	21.11%	22.51%	23.90%	22.15%

จากตารางที่ 6 แสดงให้เห็นว่าในลำดับการทดลองที่ 1 ถึงการทดลองที่ 20 เมื่อจำแนกบทเพลงไทยด้วยการวัดความคล้ายคลึงโคไซน์โดยไม่มีการลดความซ้ำซ้อนของคุณลักษณะ ให้ผลประสิทธิภาพด้านความถูกต้องในการจำแนกร้อยละ 57 โดยใช้เวลาในการประมวลผล 2.30 วินาที และในการทดลองที่ 14 เมื่อจำแนกบทเพลงไทยด้วยการวัดความคล้ายคลึงโคไซน์โดยมีการลดความซ้ำซ้อนของคุณลักษณะลงร้อยละ 70 พบว่าประสิทธิภาพในการจำแนกบทเพลงไทยมีความถูกต้องร้อยละ 75.092 โดยใช้เวลาในการประมวลผล 0.76 วินาที ซึ่งให้ประสิทธิภาพด้านความถูกต้องในการจำแนกสูงที่สุดในรอบการทดลองทั้ง 20 รอบ นอกจากนี้เมื่อผู้วิจัยได้ประเมินประสิทธิภาพความแม่นยำจากการพิสูจน์แบบไขว้ 10 รอบ พบว่า การกระจายตัวของร้อยละด้านความแม่นยำทั้ง 10 รอบ มีการกระจายตัวที่เป็นไปในทิศทางเดียวกับผลการทดลองที่ได้จากข้อมูลทดสอบโดยผลการทดลองแสดงในรูปที่ 5 และ ตารางผลการทดลองที่ 7

สรุปผล

การวิจัยครั้งนี้ได้ทำการศึกษาและนำเสนอ ขั้นตอนการลดเทอมคุณสมบัติซ้ำซ้อนสำหรับการจำแนกบทเพลงไทย ซึ่งใช้แนวทางการลดคุณสมบัติ (Feature Reduction) ของบทเพลงที่มีความซ้ำซ้อนกันด้วยค่า min (TF*IDF) จากนั้นจะพิจารณาความคล้ายคลึงของบทเพลงเพื่อสร้างโมเดลที่มีประสิทธิภาพสำหรับการจำแนกบทเพลง โดยใช้วิธีการเปรียบเทียบความคล้ายคลึงโคไซน์ (Cosine similarity) ระหว่างเวกเตอร์เสปซของสองเอกสารโดยจากการทดลองทำให้ทราบว่าการลดคุณสมบัติของบทเพลงทำให้การพยากรณ์ประเภทของบทเพลงมีความถูกต้องสูงที่สุด คือ การลดคุณสมบัติลงจำนวนร้อยละ 70 มีความถูกต้องร้อยละ 75.092 โดยใช้เวลาในการประมวลผล คือ 0.76 วินาที ซึ่งลดระยะเวลาในการประมวลผลลดลง 1.54 วินาที ทำให้เกิดประสิทธิภาพทั้งด้านความถูกต้องและด้านเวลาในการประมวลผล และเมื่อจำลองการพยากรณ์เพื่อพิสูจน์ประสิทธิภาพความแม่นยำโดยใช้การพิสูจน์แบบไขว้ 10 รอบ ผลที่ได้มีทิศทางในการพยากรณ์เป็นไปในทิศทางเดียวกัน ซึ่งผลที่ได้จากการวิจัยนี้สามารถนำไปประยุกต์ใช้กับการจำแนกเอกสารออนไลน์ได้ เช่น การวิเคราะห์เชิงอารมณ์ (Sentiment analysis) การค้นคืนเชิงอารมณ์ (Sentiment retrieval) หรือระบบการแนะนำ (Recommended system) เป็นต้น

ข้อเสนอแนะ

ในการพัฒนาขั้นตอนสำหรับการจำแนกเชิงอารมณ์ให้มีความชาญฉลาดมากยิ่งขึ้น ควรมีการใช้เทคนิคการวิเคราะห์ลำดับหน้าที่ของคำ (Part of speech: POS) เพื่อความแม่นยำมากยิ่งขึ้น โดยอาจจะใช้หลักการของคอนดิชันนอลแรนดอมฟิลด์ (Conditional Random field: CRF) ในการเรียนรู้รูปประโยคของแต่ละอารมณ์ความรู้สึก แต่ทั้งนี้การเพิ่มขั้นตอนดังกล่าวเข้ามาอาจทำให้เวลาที่ใช้ในการประมวลผลมีมากขึ้นเช่นกัน ดังนั้นจึงจำเป็นต้องหาแนวทางในการลดความซ้ำซ้อนของคุณลักษณะที่ไม่กระทบต่อการเรียนรู้ของคอนดิชันนอลแรนดอมฟิลด์ ซึ่งจะทำให้เกิดประสิทธิภาพทั้งด้านความถูกต้องและประหยัดเวลาในการประมวลผลข้อมูล นอกจากนี้ ควรเพิ่มการศึกษาเกี่ยวกับการแยกส่วนประกอบของเนื้อเพลงเพื่อวิเคราะห์ถึงความสัมพันธ์ของแต่ละส่วน โดยอาจให้ค่าน้ำหนักของแต่ละส่วนแตกต่างกันตามธรรมชาติของบทเพลง เนื่องจากส่วนต่างๆ เช่น ท่อนฮุคมีการซ้ำของเนื้อบทเพลงมากกว่าส่วนเพลงอื่น เป็นต้น ซึ่งอาจส่งผลต่อการจำแนกอารมณ์ของบทเพลงได้

เอกสารอ้างอิง

Aizawa, A. (2000). The feature quantity: An information-theoretic perspective of tfidf-likez measures. **In Proceedings of the 23rd ACM SIGIR conference on research and development in information retrieval**, 104–111.

Aizawa, A. (2001). Linguistic techniques to improve the performance of automatic text categorization. **In Proceedings of the sixth natural language processing Pacific rim symposium (NLPRS 2001)**, 307–314.

Hiemstra, D. (2000). A probabilistic justification for using tfidf term weighting in information retrieval. **International Journal on Digital Libraries**, 3(2), 131–139.

Haifeng Liu, Zheng Hu, Ahmad Mian, HuiTian, Xuzhen Zhu. (2014), A new user similarity model to improve the accuracy of collaborative filtering, **Knowledge-Based Systems**, 56, January 2014, 156-166, ISSN 0950-7051.

Kunnu, W., & Kaewrattanapat, N. (2013). The automatic classification of Thai news by similarity method. **In Proceedings of the 2th International Symposium on Business and Social Sciences**, Osaka, Japan.

Ron, K. (1995). A study of CrossValidation and Bootstrap for accuracy estimation and model selection. **In Proceeding IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2**, 1137-1143, ISBN:1-55860-363-8.