กระบวนการแนะนำการเชื่อมโยงของบทความในวิกิพีเดียภาษาไทย

นายสมภพ เซี่ยงคิ้ว

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2556
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

บทคัดย่อและแฟ้มข้อมูลฉบับเต็มของวิทยานิพนธ์ตั้งแต่ปีการศึกษา 2554 ที่ให้บริการในคลังปัญญาจุฬาฯ (CUIR)
เป็นแฟ้มข้อมูลของนิสิตเจ้าของวิทยานิพนธ์ ที่ส่งผ่านทางบัณฑิตวิทยาลัย
The abstract and full text of theses from the academic year 2011 in Chulalongkorn University Intellectual Repository (CUIR)
are the thesis authors' files submitted through the University Graduate School.

LINK SUGGESTION APPROACH FOR ARTICLES IN THAI WIKIPEDIA

Mr. Sompop Siangkhio

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2013

Copyright of Chulalongkorn University

หัวข้อวิทยานิพนธ์	กระบวนการแนะนำการเชื่อมโยงของบทความในวิกิพีเดียภาษาไทย
โดย	นายสมภพ เชี่ยงคิ้ว
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	รองศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์

(ศาสตราจารย์ ดร.บัณฑิต เอื้ออาภรณ์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.เศรษฐา ปานงาม)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

(รองศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง)

.....กรรมการภายนอกมหาวิทยาลัย

(ดร.บัณฑิต มนัสเกษมศักดิ์)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สมภพ เชี่ยงคิ้ว : กระบวนการแนะนำการเชื่อมโยงของบทความในวิกิพีเดียภาษาไทย. (LINK SUGGESTION APPROACH FOR ARTICLES IN THAI WIKIPEDIA) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร.อรรถสิทธิ์ สุรฤกษ์, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: รศ. ดร. อานนท์ รุ่งสว่าง, 37 หน้า.

วิกิพีเดียได้ถูกใช้ประโยชน์ในรูปของแหล่งความรู้ที่ถูกสร้างขึ้นโดยน้ำมือมนุษย์ในงานวิจัยหลายๆ งานด้านการประมวลผลภาษาธรรมชาติ วิทยานิพนธ์ฉบับนี้ นำเสนอการใช้เครื่องจักรเรียนรู้ และวิกิพีเดียเป็นแหล่งความรู้ สำหรับการเพิ่มความสมบูรณ์ของข้อความแบบอัตโนมัติ ขั้นตอนการทำงานของระบบที่สำคัญคือ เริ่มจากวิเคราะห์ และสกัดคำสำคัญจากบทความ และต่อมา พิจารณาเลือกหน้าวิกิพีเดียที่มีความเกี่ยวเนื่องกับคำสำคัญนั้น เพื่อแนะนำเป็นการเชื่อมโยงปลายทางไปสู่แหล่งข้อมูลเพิ่มเติม จากการทดลองในเบื้องต้นกับชุดทดสอบบทความวิกิพีเดียภาษาไทยแสดงให้เห็นว่า ระบบที่นำเสนอนี้ให้ผลลัพธ์แนะนำการเชื่อมโยงแบบอัตโตมัติได้ถูกต้องถึง 85%

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ภาควิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต .....
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....
ปีการศึกษา	2556	ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

5371450021 : MAJOR COMPUTER SCIENCE
KEYWORDS: LINK SUGGESTION / WIKIPEDIA / WEB SERVICES / ONTOLOGY

SOMPOP SIANGKHIO: LINK SUGGESTION APPROACH FOR ARTICLES IN THAI WIKIPEDIA. ADVISOR: ASST. PROF. ATHASIT SURARERKS, Ph.D., CO-ADVISOR: ASSOC. PROF. ARNON RUNGSAWANG, Ph.D., 37 pp.

Wikipedia has been used as a human engineered knowledge source for many natural language processing tasks. This thesis presents a machine learning approach and the use of Wikipedia as a knowledge source for automatic enriching a text. Given an input document, important concepts in the text have been first identified, and then chosen corresponding Wikipedia pages have been suggested as the destination links for additional information. Preliminary experiments of the system on a test set of Thai Wikipedia articles show that this automatic link suggestion approach provides reasonably up to 85% link suggestion accuracy.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Department:	Computer Engineering	Student's Signature .....
Field of Study:	Computer Science	Advisor's Signature .....
Academic Year:	2013	Co-Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จเรียบร้อยได้ด้วยดีเพราะได้รับคำแนะนำ และ ให้คำปรึกษาจาก ผู้ช่วยศาสตราจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์ และรองศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง ซึ่งเป็นผู้ชี้แนะแนวทางในการศึกษาและให้ข้อเสนอแนะที่เป็นประโยชน์ยิ่งต่อการวิจัย จนเกิดเป็นวิทยานิพนธ์ฉบับนี้ขึ้นนอกจากนี้ขอขอบคุณคณะกรรมการสอบวิทยานิพนธ์ ซึ่งได้แก่ ผู้ช่วยศาสตราจารย์ ดร.เศรษฐา ปานงาม ผู้ซึ่งเป็นประธาน ดร.บัณฑิต มนัสเกษมศักดิ์ ผู้ซึ่งเป็นกรรมการ ที่ได้สละเวลาอันมีค่ามาใช้ในการชี้แจงถึงข้อบกพร่อง รวมถึงแนวทางการแก้ไขและข้อแนะนำดีๆ ให้แก่ข้าพเจ้า

ขอขอบพระคุณ พี่ๆเพื่อนๆ น้องๆที่คอยให้คำแนะนำโดยเฉพาะสมาชิกห้องปฏิบัติการวิจัย MIKE LAB และสมาชิกห้องปฏิบัติการวิจัย ELITE LAB เมื่อข้าพเจ้าพบปัญหาต่างๆ ในขณะทำงานวิจัยเล่มนี้ทำให้วิทยานิพนธ์เล่มนี้ ทำให้สำเร็จสมบูรณ์ไปด้วยดี และยิ่งไปกว่านั้นขอขอบคุณภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ที่ได้กรุณาอนุญาตให้ใช้สถานที่ ทรัพยากร และเครื่องมือต่างๆ สำหรับทำวิจัยและขอขอบคุณต่อเจ้าหน้าที่ของภาควิศวกรรมศาสตร์คอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ทุกท่านที่อำนวยความสะดวกในทุกๆ เรื่องที่เกี่ยวกับการศึกษา

ขอกราบขอบพระคุณคุณพ่อ คุณแม่ และสมาชิกในครอบครัวที่ให้ความรัก ความห่วงใย และเป็นกำลังใจแก่ผู้วิจัยในการดำเนินชีวิตมาโดยตลอด

ขอบคุณเพื่อนๆ ที่มีส่วนช่วยในการติดต่อประสานงานให้งานวิจัยสำเร็จลุล่วง พร้อมด้วยเพื่อน ๆ พี่ ๆ น้อง ๆ ทุกคนที่เป็นกำลังใจให้กันเสมอมา ที่ให้คำแนะนำ ความช่วยเหลือต่างๆ ตลอดระยะเวลาที่ดำเนินการวิจัย

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญ

หน้า

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

หน้า

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญภาพ

หน้า

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญตาราง

หน้า

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

วิกิพีเดีย (Wikipedia) เป็นสารานุกรมที่ถูกกล่าวถึง และใช้งานมากที่สุด อีกทั้งยังเป็นแหล่งความรู้แบบออนไลน์ที่ใหญ่ที่สุดบนโลกของอินเทอร์เน็ต รวมทั้งมีบทความมากมายในหลายๆ ภาษา [1] วลี หรือคำที่สำคัญ (important keyword) ในบทความวิกิพีเดียจะถูกระบุ และกำหนดลิงค์เชื่อมโยง (link) ไปยังบทความปลายทางอื่นๆ เพื่อให้ผู้ใช้เข้าถึงข้อมูลเพิ่มเติมที่เกี่ยวข้องได้อย่างรวดเร็วขึ้น สืบเนื่องจากทั้งการเลือกคำสำคัญ และการเลือกบทความปลายทางที่เกี่ยวข้อง ได้ถูกคัดเลือก และกำหนดโดยผู้เขียนบทความวิกิพีเดีย (Wikipedia contributors) นักวิจัยส่วนหนึ่งจึงได้ทำการวิจัยเพื่อสกัดเอาความรู้จากวิกิพีเดียเพื่อไปประยุกต์ใช้ในกระบวนการทำงานด้านการประมวลผลภาษาธรรมชาติ (natural language processing) [2],[3],[4],[5] ในบทความนี้เราจะสนใจในกรรมวิธีการแนะนำการเชื่อมโยงไปยังบทความวิกิพีเดียภาษาไทย ซึ่งประกอบไปด้วยขั้นตอนการระบุคำสำคัญจากบทความต้นทาง และขั้นตอนการเลือกบทความวิกิพีเดียปลายทางที่มีความเกี่ยวข้องกับคำสำคัญนั้นๆ แบบอัตโนมัติ

ปัจจุบัน วิกิพีเดียภาษาไทยประกอบด้วยบทความทั้งหมดประมาณ 80,000 บทความ [1] ซึ่งจำนวนบทความภาษาไทยนี้มีขนาดจำนวนเป็นเพียงหนึ่งในห้าส่วนเมื่อเทียบกับบทความวิกิพีเดียภาษาอังกฤษ ดังนั้น เพื่อเป็นการส่งเสริม กระตุ้น และอำนวยความสะดวกให้กับผู้ใช้งานวิกิพีเดียให้สามารถสร้างบทความวิกิพีเดียภาษาไทยได้ง่ายยิ่งขึ้น ระบบแนะนำการเชื่อมโยง (link suggestion) แบบอัตโนมัติจึงเป็นเรื่องที่จำเป็น จากที่เราศึกษามานั้น พบว่า ณ เวลานี้ งานวิจัยชิ้นนี้ สามารถถูกจัดได้ว่าเป็นงานชิ้นแรกๆ ที่พยายามศึกษาเรียนรู้ขั้นตอนวิธีการสร้างการเชื่อมโยงภายในจากวิกิพีเดียภาษาไทย และเนื่องจากว่าภาษาไทยนั้นมีลักษณะการเขียนที่แตกต่างจากภาษาอังกฤษมาก ตัวอย่างเช่น ไม่มีช่องว่างแยกระหว่างคำที่ชัดเจน ไม่มีการแปลงรูปกริยาเพื่อแสดงเวลา (tense) ไม่มีการใช้อักษรตัวใหญ่เพื่อระบุว่าเป็นคำนามเฉพาะ (proper noun) เป็นต้น ดังนั้นทุกวิธีการที่ถูกนำเสนอในงานวิจัยที่ผ่านมา [3],[4] ที่สามารถทำงานได้ผลดีกับวิกิพีเดียภาษาอังกฤษ จึงไม่อาจจะให้ผลลัพธ์การทำงานที่มีประสิทธิภาพในการสร้างการเชื่อมโยงเมื่อถูกนำมาใช้กับข้อความภาษาไทยได้ ในงานวิจัยนี้เราจึงนำเสนอ กระบวนการแนะนำการเชื่อมโยงแบบอัตโนมัติ โดยเฉพาะสำหรับวิกิพีเดียภาษาไทย

ระบบแนะนำการเชื่อมโยงวิกิพีเดียภาษาไทยที่ถูกนำเสนอนี้ ได้รวบรวมเอาสองกระบวนการทำงาน กล่าวคือ ขั้นตอนการสกัดคำสำคัญ (keyword extraction) และขั้นตอนการแก้ปัญหาลิงก์ที่มีความกำกวม (link disambiguation) เข้าไว้ในกรอบการทำงานเดียวกัน โดยเริ่มต้นจากข้อมูลนำเข้าที่เป็นเอกสาร ระบบจะสกัดคุณลักษณะ (feature) จากคำสำคัญที่พ้องตรงกับคำในพจนานุกรมคำศัพท์ควบคุม (controlled vocabulary) ที่ได้สร้างไว้ก่อนหน้า แล้วจึงทำการคำนวณ คุณลักษณะของคำสำคัญที่สกัดได้ให้กับเครื่องจักรเรียนรู้เลือกว่าคำสำคัญนั้นควรจะถูกสร้างลิงก์ ณ ตำแหน่งที่

ปรากฏอยู่ในเอกสารนั้น ในกรณีที่คำสำคัญที่กำลังถูกพิจารณาอยู่เป็นคำที่มีความหมายกำกวม (ambiguous word) ระบบจะทำการตรวจสอบความคล้ายคลึงตามความหมายของคำบริบท (context words) รอบๆ คำกำกวมกับบทความปลายทาง เพื่อเลือกบทความปลายทางที่เหมาะสมต่อไป ผลการทดลองเบื้องต้น ที่ได้จากการใช้ชุดข้อมูลจากวิกิพีเดียภาษาไทยของเดือนกรกฎาคม 2555 ได้แสดงให้เห็นว่า วิธีการที่นำเสนอในบทความนี้ สามารถทำงานได้อย่างมีประสิทธิภาพ เป็นที่น่าพอใจ

ในย่อหน้าต่อๆ ไป เราจะเริ่มต้นด้วยการพูดถึงงานวิจัยที่เกี่ยวข้องกับวิธีการสร้างการเชื่อมโยงของวิกิพีเดียภาษาอังกฤษโดยพอสังเขป หลังจากนั้น เราจะนำเสนอภาพรวมของระบบแนะนำการเชื่อมโยงวิกิพีเดียภาษาไทย โดยเริ่มอธิบายตั้งแต่วิธีการเตรียมข้อมูลวิกิพีเดียภาษาไทย การสกัดคุณลักษณะ การฝึกสอน และทดสอบเครื่องจักรเรียนรู้ หลังจากนั้น เราจะได้กล่าวถึงประสิทธิภาพของระบบจากผลการทดลองในเบื้องต้นที่ได้ และสิ้นสุดบทความนี้ด้วยสรุป และแนวทางทำวิจัยต่อเพิ่มเติมในอนาคต

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1.2 วัตถุประสงค์ของการวิจัย

1. นำเสนอกระบวนการในการสร้างการเชื่อมโยงกันระหว่างบทความต่างๆ แบบอัตโนมัติอย่างเหมาะสม สำหรับวิกิพีเดียภาษาไทย

1.3 ขอบเขตของการวิจัย

1. ข้อมูลที่นำมาใช้ในงานวิจัย เป็นข้อมูลจากวิกิพีเดียภาษาไทย “www.th.wikipedia.org”
2. การแบ่งกลุ่มข้อมูล จะใช้โปรแกรมสำเร็จรูป Weka [6]
3. การตัดคำภาษาไทย จะใช้โปรแกรมสำเร็จรูป Lexto [7]
4. สร้างโมเดลในการแนะนำการเชื่อมโยงของบทความแบบอัตโนมัติ โดยอาศัยคุณลักษณะเฉพาะของบทความและ ความสัมพันธ์กันของบทความต่างๆ

1.4 ประโยชน์ที่ได้รับ

1. ได้ระบบแนะนำการเชื่อมโยงแบบอัตโนมัติเพื่ออำนวยความสะดวกให้กับผู้สร้างบทความในวิกิพีเดียภาษาไทย ทำให้บทความมีความน่าสนใจมากขึ้น

1.5 ลำดับการจัดเรียงเนื้อหาในวิทยานิพนธ์

วิทยานิพนธ์นี้แบ่งเนื้อหาออกเป็น 6 บทดังต่อไปนี้ บทที่ 1 เป็นบทนำซึ่งกล่าวถึง ความเป็นมาและความสำคัญของปัญหา รวมถึงวัตถุประสงค์ของการวิจัย บทที่ 2 กล่าวถึงทฤษฎีพื้นฐานและงานวิจัยที่เกี่ยวข้องในงานวิจัยนี้ บทที่ 3 กล่าวถึงแนวคิดและวิธีการดำเนินงานวิจัย บทที่ 4 กล่าวถึงการออกแบบและพัฒนาระบบ บทที่ 5 กล่าวถึงวิธีการทดสอบระบบ และบทที่ 6 กล่าวถึงสรุปผลการวิจัยและข้อเสนอแนะ

1.6 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้รับการตีพิมพ์เป็นผลงานวิชาการในหัวข้อเรื่องดังต่อไปนี้

1. "Thai Wikipedia Link Suggestion Framework" โดย อานนท์ รุ่งสว่างบัณฑิต มนัสเกษมศักดิ์, สมภพ เชี่ยงคิ้ว, อรรถสิทธิ์ สุรฤกษ์ และ บัณฑิต มนัสเกษมศักดิ์ ในงานประชุมวิชาการ The 5th FTRA International Conference on Information Technology Convergence and Service (ITCS-13)

2. "Thai Wikipedia Link Suggestion Approach" โดย สมภพ เซี่ยงคิ้ว, อรรถสิทธิ์ สุรฤกษ์, บัณฑิต มนัสเกษมศักดิ์ และ อานนท์ รุ่งสว่าง ในงานประชุมวิชาการ The 10th International Joint Conference on Computer Science and Software Engineering (JCSSE13)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การสร้างลิงก์ในบทความวิกิพีเดีย

การสร้างลิงก์ในบทความวิกิพีเดีย (Wikification) [8] คือ ข้อความภายในบทความวิกิพีเดียคำสำคัญที่จะถูกสร้างลิงค์จะต้องมี เครื่องหมายวงเล็บก้ามปู ([[...]]) ข้อความภายในเครื่องหมายจะถูกสร้างเป็นลิงค์ไปยังบทความปลายทาง ซึ่งขึ้นอยู่กับผู้เขียนบทความนั้นว่าจะต้องการสร้างลิงค์ไปยังบทความใด ซึ่งจะทำเพียงครั้งเดียวในคำแรกที่ปรากฏในบทความ ถึงแม้ว่าคำสำคัญนั้นปรากฏในบทความมากกว่าหนึ่งครั้งก็ตาม ชนิดของหน้าบทความวิกิพีเดียที่ผู้ใช้พบระหว่างการค้นหาข้อมูลได้แก่

1. บทความทั่วไป

   หน้าบทความทั่วไป คือ หน้าบทความที่ประกอบด้วยเนื้อหาต่างๆ ที่เกี่ยวข้องกับหัวข้อเรื่องของบทความนั้น อาจมีโครงแบบ (Template) ที่แตกต่างกันออกไป เช่น ชื่อบทความ จุฬาลงกรณ์มหาวิทยาลัย เป็นต้น ดังภาพที่ 2.1

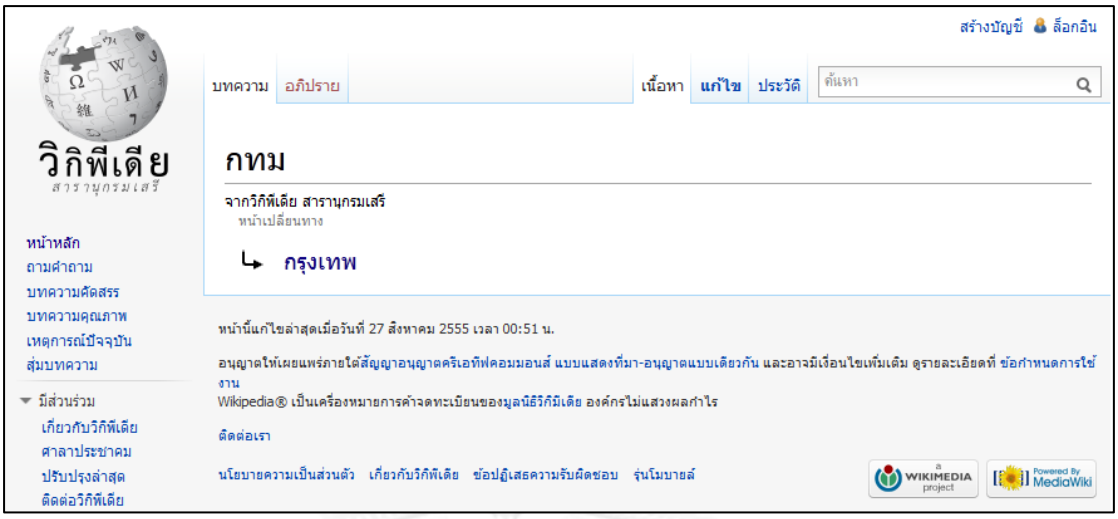
ภาพที่ 2.1 ตัวอย่างของบทความวิกิพีเดีย

2. บทความที่เป็นหน้ารายการ

   หน้าบทความที่ทำหน้าที่เหมือนเป็นจุดรวมของรายการบทความ (Hub) เช่น ชื่อบทความ รายชื่อจังหวัดในประเทศไทยเรียงตามพื้นที่ เป็นต้น

3. บทความที่เป็นหน้าเปลี่ยนทาง (Redirect page)

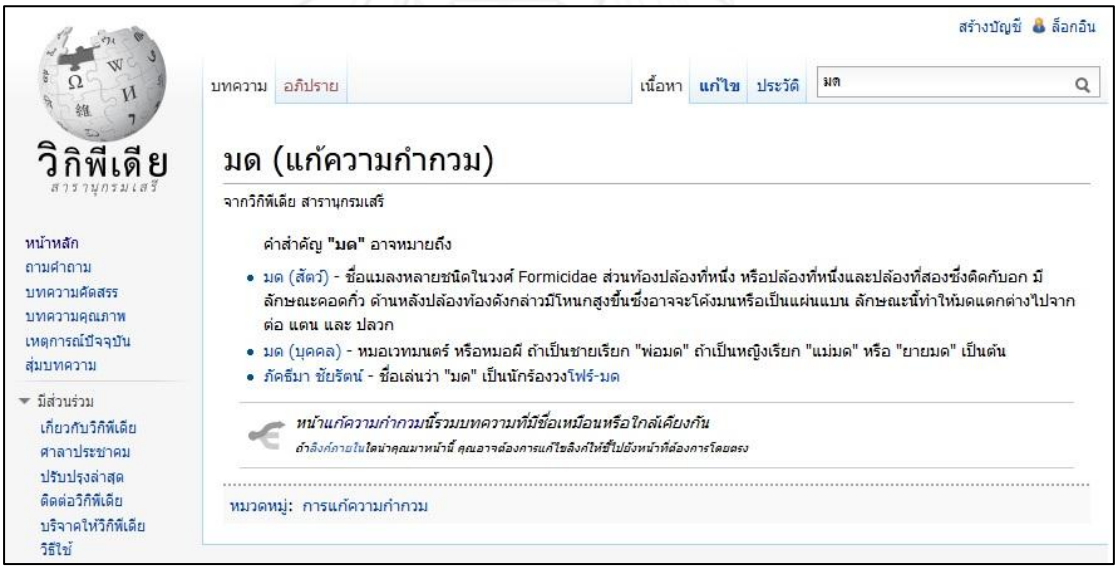
   หน้าบทความที่ผู้ใช้เข้ามาแล้วจะถูกเปลี่ยนทางไปหน้าบทความต้นฉบับที่เกี่ยวข้องแทน เช่น ชื่อบทความ กทม (จะถูกเปลี่ยนทางไปยังหน้า กรุงเทพมหานคร) , USB (จะถูกเปลี่ยนทางไปยังหน้า ยูเอสบี) ดังภาพที่ 2.2



ภาพที่ 2.2 ตัวอย่างของบทความเปลี่ยนทาง

4. บทความที่เป็นหน้าแก้ไขคำกำกวม (Disambiguation page)

   หน้าบทความที่ใช้สำหรับชื่อบทความที่มีความกำกวม อาจมีหลายนัย ซึ่งหน้าบทความนี้จะเป็นรายการของบทความปลายทางที่เกี่ยวข้องกับชื่อบทความนั้น เช่น ชื่อบทความ มด ซึ่งอาจหมายถึง มด (แมลง) หรือ มด (บุคคล) ดังภาพที่ 2.3

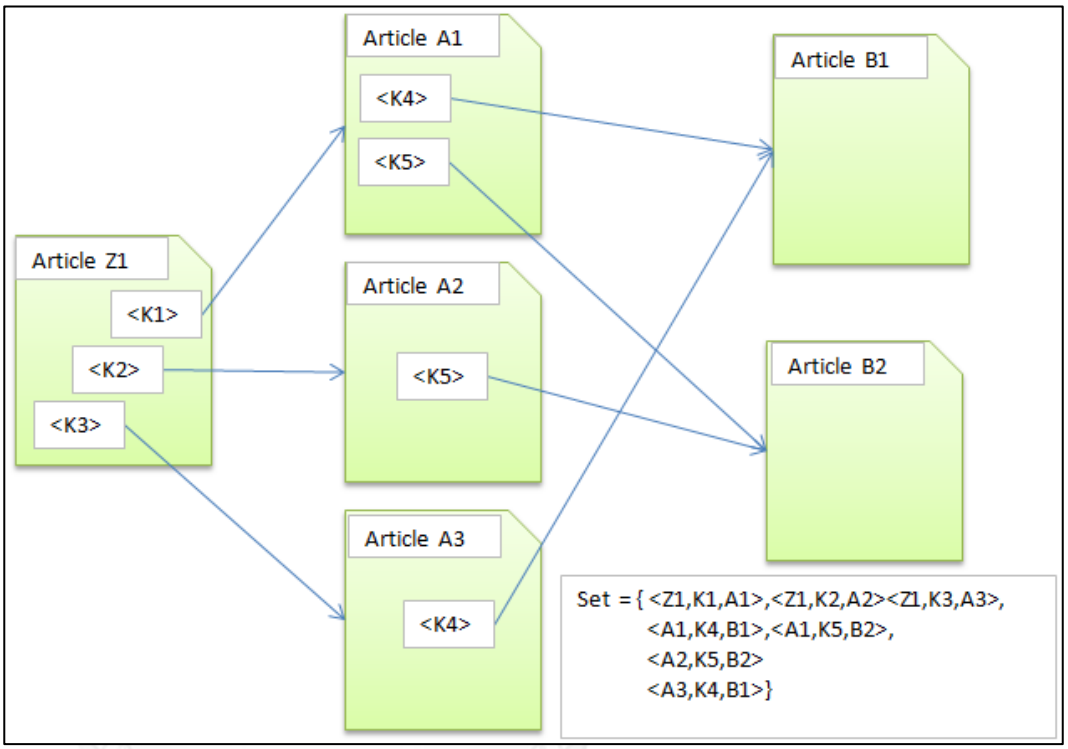


ภาพที่ 2.3 ตัวอย่างของบทความแก้ไขคำกำกวม

2.1.2 การวิเคราะห์ลิงก์

การวิเคราะห์ลิงก์ (Link analysis) คือ เราสามารถมองได้ว่าบทความในวิกิพีเดียนั้น มีการเชื่อมโยงกันในรูปแบบกราฟ ดังภาพที่ 2.3 โดยบทความจะแทนด้วยโหนด (vertex) และการเชื่อมโยงด้วยคำสำคัญในบทความหนึ่งไปยังบทความอื่นๆ ในวิกิพีเดียนั้น เสมือนเส้นการเชื่อมโยงระหว่างโหนด (edge) การวิเคราะห์ลิงก์ทำให้ทราบถึงการเชื่อมโยงเข้าหากันแต่ละบทความ คำสำคัญที่ใช้ในการเชื่อมโยง ซึ่งสามารถนำมาใช้ประโยชน์ในงานวิจัยนี้ได้ โดยอาจวิเคราะห์จาก In-degree คือ

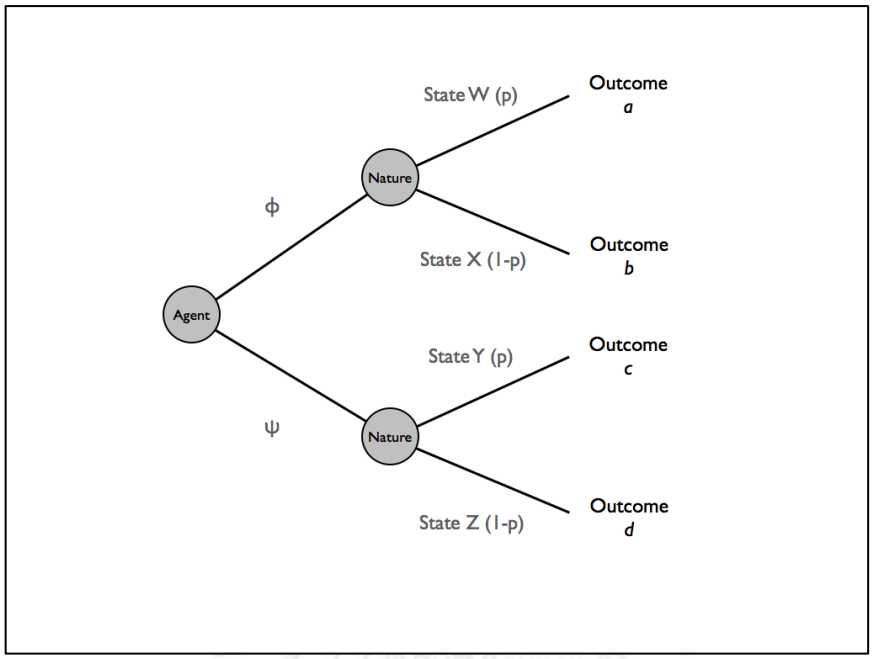
จำนวนลิงก์หรือการเชื่อมโยงจากบทความอื่นๆ มายังบทความที่สนใจ หาก In-link มีจำนวนมาก แสดงให้เห็นว่าบทความอื่นๆ จำนวนมาก ชี้มายังบทความที่สนใจนี้มากเช่นเดียวกัน เพราะการทำลิงก์ในบทความหนึ่งๆ ที่ถูกต้องและเหมาะสม จะทำเพียงแค่ครั้งเดียว คือครั้งแรกที่ปรากฏคำที่ต้องการอ้างอิงเท่านั้น Out-degree คือ จำนวนของลิงก์หรือการเชื่อมโยงจากบทความที่เราสนใจไปยังบทความอื่นๆ ซึ่งค่าสถิติเหล่านี้สามารถนำไปคำนวณเพื่อเป็นมาตรวัดเพื่อบ่งบอกความสำคัญของเอกสารโดยใช้อัลกอริทึม เช่น อัลกอริทึม HITS [9] หรือ เพจแรงค์ (Page Rank) [10] ได้เป็นต้น



ภาพที่ 2.4 ตัวอย่างการเชื่อมโยงของวิกิพีเดีย

2.1.3 ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ (Decision Tree) เป็นเครื่องจักรเรียนรู้ซึ่งใช้ในการตัดสินใจ ต้นไม้ตัดสินใจจะถูกแบ่งกิ่งด้วยคุณลักษณะที่แตกต่างกัน และโหนดใบจะเป็นผลการตัดสินใจ การสร้างโมเดลต้นไม้ตัดสินใจสำคัญที่การแบ่งกิ่งเพื่อให้ได้ต้นไม้ที่นำไปใช้จริงได้ความถูกต้องสูงที่สุด หลักเกณฑ์ว่าจะแบ่งจากคุณลักษณะใดก่อนหรือหลังนั้นอาจพิจารณาจากค่า GINI, Information Gain และ Gain Ratio [11] เป็นต้น เมื่อได้ต้นไม้ตัดสินใจแล้วการทำนายคำตอบ จะเริ่มจากโหนดรากของต้นไม้ และพิจารณาค่าของคุณลักษณะต่างๆ และโหนดใบจะเป็นผลลัพธ์จากการทำนาย ตัวอย่างของต้นไม้ตัดสินใจ



ภาพที่ 2.5 ตัวอย่างต้นไม้ตัดสินใจ

2.1.4 ตัวจำแนกเบย์อย่างง่าย

ตัวจำแนกเบย์อย่างง่าย (Naïve Bayes) เป็นเครื่องจักรเรียนรู้ที่อาศัยหลักการความน่าจะเป็น (Probability) ตามทฤษฎีของเบย์ (Bayes theorem) ซึ่งมีอัลกอริทึมที่ไม่ซับซ้อน โดยพิจารณาจากค่าความน่าจะเป็นของแต่ละคุณลักษณะสำหรับคลาสคำตอบที่แตกต่างกัน โดยมีสมมติฐานของการเป็นอิสระต่อกันของคุณลักษณะที่นำมาใช้และเหมาะสำหรับการจำแนกที่มีจำนวนคุณลักษณะค่อนข้างมาก คลาสคำตอบที่ได้จะเป็นคลาสที่สามารถคำนวณค่าความน่าจะเป็นภายหลังจากเงื่อนไขและคุณลักษณะสูงที่สุด

2.1.5 ตัวจำแนกซัพพอร์ทเวกเตอร์แมชชีน

ตัวจำแนกซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine - SVM) เป็นเครื่องจักรเรียนรู้อีกเทคนิคหนึ่งของศาสตร์ด้านเครื่องจักรเรียนรู้ (Machine Learning) ซึ่งนำมาใช้ทั้งปัญหาการจำแนก (Classification) และ การถดถอย (Regression) โดยตัวจำแนก SVM โดยปกติจะเป็น Binary classifier คือสามารถจำแนกข้อมูลได้ออกเป็นสองกลุ่มเท่านั้น ตัวอย่างดังรูปที่ 3 ใช้ระนาบพยายามแบ่งข้อมูลออกเป็นสองส่วนให้ได้ระยะห่างของขอบการแบ่งสูงที่สุด และยังสามารถเลือกเคอร์เนลฟังก์ชัน ใช้สาหรับเปลี่ยนระนาบขอบคุณลักษณะเดิม ไปยังระนาบคุณลักษณะใหม่ที่แตกต่างกันออกไป อาจทำให้การจำแนกคุณลักษณะที่มีอยู่เดิมนั้นง่ายขึ้น และอาจจำแนกได้ถูกต้องมากยิ่งขึ้น

2.1.6 การวัดค่าความเหมือนระหว่างสองเอกสาร

การวัดค่าความเหมือนของเอกสารทำได้โดยการเปลี่ยนเอกสารที่ต้องการนั้นให้อยู่ในรูปของ Vector Space Model (VSM) ซึ่งเป็นวิธีการหนึ่งสำหรับสร้างตัวแทนตัวเอกสาร (Document representation) สำหรับเอกสารภาษาไทยจะต้องผ่านกระบวนการตัดคำก่อน เพื่อได้เป็นคำ (term)

และนำคำเหล่านั้นมาแสดงในรูปของเวกเตอร์ และค่าในแต่ละมิติของเวกเตอร์นั้นอาจให้น้ำหนักที่แตกต่างกันตามความถี่ (TF) หรือค่า TF-IDF [12] เป็นต้น การแสดงตัวแทนเอกสารลักษณะนี้อาจมองเป็นลักษณะของถุงของคำ (bag of words) นั่นคือคำต่างๆ ไม่ขึ้นต่อกัน การวัดค่าความสัมพันธ์ของเอกสาร *A* และ *B* ซึ่งพิจารณาจากสมการต่อไปนี้

$$\boldsymbol{Similarity = Cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}} \quad (1)$$

เมื่อ A แทนเวกเตอร์ ของเอกสารต้นทาง

B แทนเวกเตอร์ ของเอกสารปลายทาง

ค่าความเหมือนของเอกสารที่คำนวนได้จากสมการนี้มีค่าระหว่าง 0 ถึง 1 ยิ่งค่าความเหมือนของเอกสารเข้าใกล้ 1 แสดงว่าเอกสารสองเอกสารนั้นมีความสัมพันธ์หรือคล้ายคลึงกันมาก

2.1.7 การตัดคำและการแปลงรูปคำ

เนื่องจากภาษาไทยมีลักษณะที่เขียนตัวอักขระติดกันทั้งหมดไม่มีเครื่องหมายวรรคตอนขั้นระหว่างคำ หากต้องการให้คอมพิวเตอร์ประมวลผลทางภาษา (Language processing) จำเป็นต้องแบ่งคำออกเป็นส่วนๆ เพื่อแสดงถึงหน่วยของคำได้ โดยทั่วไปแล้วการตัดคำภาษาไทยมีวิธีการที่ใช้ในปัจจุบันเป็นหลักอยู่ 3 วิธี ได้แก่

1. การใช้กฎ

   การสร้างพยางค์ไทยที่ประกอบด้วยพยัญชนะ สระ วรรณยุกต์ ตัวสะกด ตัวการันต์ แนวทางนี้ทำได้ง่ายที่สุด ทำงานได้เร็วที่สุด แต่แบ่งคำพยางค์เดียวได้เท่านั้น ไม่สามารถจัดการกับคำหลายพยางค์ได้

2. การใช้พจนานุกรม

   ต้องทำรายการคำเอาไว้ล่วงหน้า เมื่อต้องการแบ่งคำก็เปรียบเทียบข้อความที่ต้องการแบ่งกับรายการคำที่เก็บ ไว้ในพจนานุกรม อาจพิจารณาตัดคำแบบยาวสุด คือการตัดคำที่เป็นไปได้ให้ได้คำยาวที่สุด หรือตัดคำให้ได้จำนวนคำน้อยสุด คือการตัดคำโดยการเลือกรูปแบบการตัดที่จำนวนคำที่ตัดออกมาได้มีจำนวนน้อยที่สุด

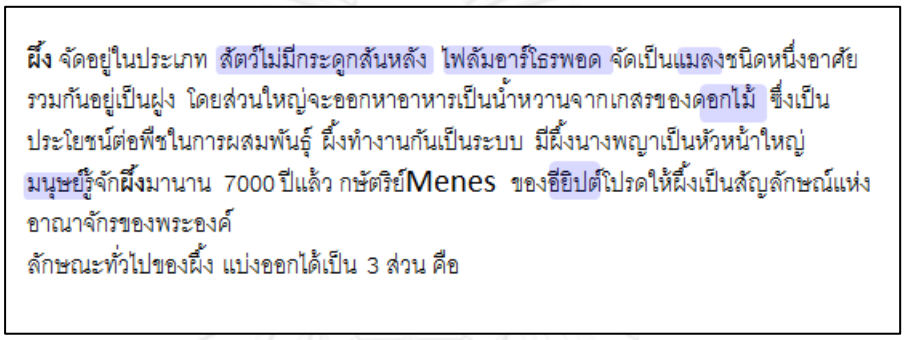
3. การใช้เทคนิดการเรียนรู้ด้วยเครื่อง

   เป็นวิธีที่ได้รับความนิยมที่สุดในปัจจุบัน โดยการฝึกฝนระบบด้วยคลังข้อความขนาดใหญ่ที่มีการแบ่งคำไว้อย่างถูกต้องแล้ว เพื่อให้เครื่องได้เรียนรู้ด้วยตนเอง จากการเก็บสถิติ

การแปลงรูปแบบคำ เป็นการลดรูปแบบที่หลากหลาย (Word variation) ของคำศัพท์ลงให้อยู่ในรูปแบบเดียวกัน ทาให้การตัดคำมีประสิทธิภาพและสามารถระบุว่าคำใดเหมือนหรือแตกต่างกันได้ ในเอกสารภาษาไทยที่พบจริงคำที่มีความหมายเหมือนกันหรือคำเดียวกัน อาจเขียนได้หลายรูปแบบ กระบวนการที่นามาใช้ ได้แก่ การกำจัดช่องว่าง (whitespace) การกำจัดอักขระพิเศษ การทำให้เป็นตัวอักษรพิมพ์เล็กสำหรับภาษาอังกฤษ

2.1.8 การสกัดคำสำคัญ

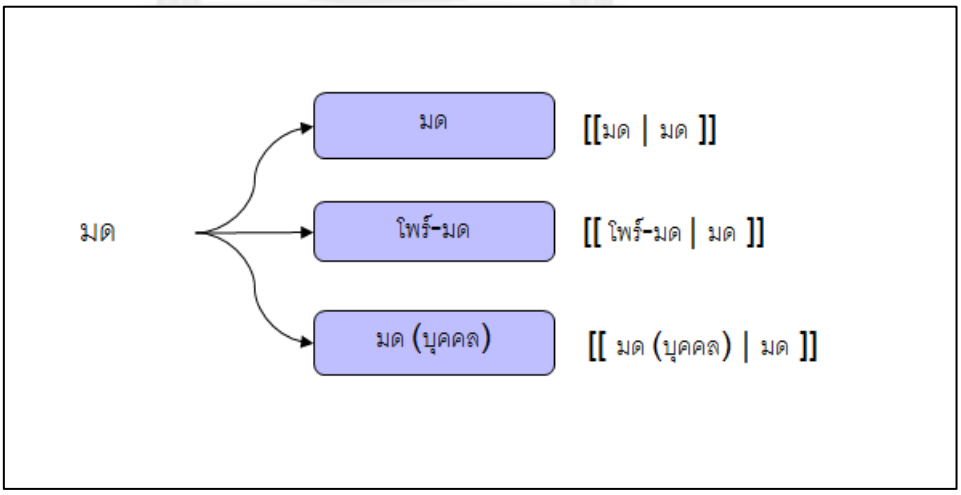
การสกัดคำสำคัญ (Keyword extraction) คือ การเลือกคำที่ปรากฏในบทความหรือเอกสารใดๆก็ตามเพื่อระบุให้คำนั้นเป็นคำสำคัญควรที่จะถูกสร้างการเชื่อมโยง ดังภาพที่ 2.6



ภาพที่ 2.6 ตัวอย่างการสกัดคำสำคัญ

2.1.9 ปัญหาความกำกวม

ปัญหาความกำกวม (Disambiguation) คือ เลือกความหมายที่เหมาะสมให้กับคำที่มีความหมายได้หลายเมื่อเขียนด้วยรูปแบบเดียวกันในที่นี้เราจะแสดงตัวอย่าง ดังภาพที่ 2.7



ภาพที่ 2.7 ตัวอย่างการปัญหาความกำกวม

2.2 งานวิจัยที่เกี่ยวข้อง

ปัญหาการแนะนำการเชื่อมโยงให้กับบทความวิกิพีเดีย ได้ถูกศึกษา และวิจัย โดยนักวิจัยในต่างประเทศหลายๆ ท่าน ผ่านการใช้บทความวิกิพีเดียภาษาอังกฤษ นักวิจัยส่วนใหญ่ได้แบ่งกระบวนการออกเป็นสองขั้นตอน [2],[3],[4],[5] กล่าวคือ ขั้นตอนการสกัดคำสำคัญ และขั้นตอนการแก้ไขปัญหาคำกำกวม (word sense disambiguation) ในที่นี้ เราจะเพียงกล่าวถึงงานวิจัยของ Mihalcea และ Csomai [3] และงานวิจัยของ Milne และ Witten [4] เท่านั้น เนื่องจากว่างานวิจัยทั้งสองเรื่องนี้ ถือได้ว่าเป็นต้นแบบของงานวิจัยด้านการแนะนำการเชื่อมโยงที่เกิดขึ้นในภายหลัง

2.2.1 งานวิจัย Wikify! Linking Documents to Encyclopedic Knowledge

Mihalcea และ Csomai [3] ได้เรียกระบบที่ตนนำเสนอว่า Wikify และเป็นทีมนักวิจัยในยุคแรกๆ ที่ให้ชื่อกรรมวิธีการแนะนำการเชื่อมโยงในลักษณะนี้ว่าเป็นการทำ wikification ระบบดังกล่าวได้แบ่งกระบวนการทำงานเป็นสองขั้นตอน การสกัดคำสำคัญ ได้แก่การระบุคำ หรือวลีที่สำคัญ เพื่อแนะนำต่อไปว่าควรที่จะสร้างการเชื่อมโยงหรือไม่ ซึ่งในขั้นตอนนี้จะประกอบไปด้วยอีกสองขั้นตอนย่อย ได้แก่ การสกัดแคนดิเดต (candidate extraction) และจัดลำดับคำสำคัญ (keyword ranking) การสกัดแคนดิเดต คือการวิเคราะห์เอกสาร และสกัดเอ็นแกรม (N-grams) ที่เป็นไปได้ทั้งหมดเมื่อเทียบกับชื่อบทความวิกิพีเดีย และแองเคอร์ (anchor text) ที่มนุษย์ได้สร้างการเชื่อมโยงไว้แล้วผ่านข้อความนั้นๆ หลังจากนั้น ขั้นตอนการจัดลำดับคำสำคัญจะทำการจัดลำดับให้กับคำ หรือวลีที่พบจากขั้นตอนการสกัดแคนดิเดต โดยการคำนวณค่าคะแนนความน่าจะเป็นในการสร้างการเชื่อมโยง (link probability) ให้กับแต่ละแคนดิเดต และทำการจัดลำดับจากค่าคะแนนดังกล่าว สุดท้ายระบบจะเลือกแคนดิเคตที่มีค่าคะแนนความน่าจะเป็นในการสร้างการเชื่อมโยงที่เกินกว่าเกณฑ์ที่กำหนด (threshold) ไว้ เป็นคำแคนดิเดตคำสำคัญเพื่อพร้อมที่จะถูกสร้างการเชื่อมโยงไปยังบทความวิกิพีเดียปลายทางในขั้นตอนต่อไป

ในกรณีที่แคนดิเดตคำสำคัญที่ถูกเลือกไว้ มีความกำกวม (ambiguity) เช่นคำดังกล่าวนั้นไม่ได้ถูกเขียนให้ตรงกับชื่อ (title) ของเอกสารวิกิพีเดีย หรือเป็นคำที่มีความหมายมากกว่าหนึ่งความหมาย การจะเลือกบทความปลายทางให้ตรงกับความหมายนั้นจะต้องเลือกโดยดูที่เนื้อหาที่ปรากฏรอบๆ (surrounding context) แคนดิเดตคำสำคัญนั้น ซึ่งวิธีการที่ Wikify เลือกใช้มีอยู่ด้วยกันสองวิธีการ [3] กล่าวคือ วิธีการแรก ระบบพยายามที่จะตีความความหมายของคำกำกวมด้วยการคำนวณความเหมือนของคำที่ปรากฏรอบๆ คำกำกวม เทียบกับบทความวิกิพีเดียปลายทางที่เป็นไปได้สำหรับคำกำกวมที่กำลังพิจารณาอยู่นั้น วิธีการที่สอง ระบบจะสกัดคุณลักษณะจากของทั้งแคนดิเดตคำสำคัญ และของทั้งคำอื่นๆ ที่ปรากฏอยู่รอบๆ แคนดิเดตคำสำคัญนั้นว่ามีหน้าที่อย่างไรในประโยค (part of speech) และนำมาเปรียบเทียบกับคุณลักษณะที่สกัดได้ในทำนองเดียวกันจากข้อมูลฝึกสอน

2.2.2 งานวิจัย Learning to Link with Wikipedia

Milne และ Witten [4] ได้นำเสนอวิธีการใช้เครื่องจักรเรียนรู้เพื่อแก้ไขปัญหาคำกำกวมเป็นขั้นตอนแรก ก่อนที่จะทำการสกัดแคนดิเดตคำสำคัญในขั้นตอนที่สอง พวกเขาจะเลือกบทความวิกิพีเดียที่มีการเชื่อมโยง หรือลิงค์จำนวนมากนำมาเป็นชุดข้อมูลฝึกสอน และกำหนดให้บทความวิกิพีเดียที่มีการทำลิงก์จากแองเคอร์เป็นชุดตัวอย่างบวก(positive example) โดยให้บทความที่เหลือเป็นชุดตัวอย่างลบ (negative example) ในแต่ละตัวอย่างฝึกสอนจะสกัดคุณลักษณะที่สำคัญอยู่สองคุณลักษณะ กล่าวคือ ค่าความน่าจะเป็นการถูกทำลิงค์ (commonness) และค่าความคล้ายคลึงกันของเอกสาร (relatedness) โดยที่ค่าความน่าจะเป็นของการถูกทำลิงค์ คือจำนวนครั้งที่บทความถูกเลือกเป็นบทความปลายทางในฐานข้อมูลวิกิพีเดีย ส่วนค่าความคล้ายคลึงกันของบทความจะถูกเปรียบเทียบกันระหว่างบทความที่ถูกเลือกเป็นบทความปลายทางกับคำที่ไม่กำกวมที่ปรากฏรอบๆ คำกำกวมที่กำลังถูกพิจารณานั้น

ในทางตรงกันข้ามวิธีการสกัดคำสำคัญของระบบ Wikify [3] ที่อาศัยเฉพาะค่าความน่าจะเป็นในการทำลิงค์ Milne และ Witten [4] ได้นำวิธีการเครื่องจักรเรียนรู้มาใช้ โดยสกัดคุณลักษณะจากคำสำคัญเพิ่มเติมจากค่าความน่าจะเป็นของถูกทำลิงค์ และค่าความคล้ายคลึงกันของเอกสารที่ผ่านมา ซึ่งได้แก่ เปอร์เซ็นต์ค่าความมั่นใจในการแก้ไขปัญหาความกำกวม (disambiguation confidence) ซึ่งจะได้จากส่วนของการแก้ไขปัญหาความกำกวมในขั้นตอนแรกที่ผ่านมา ค่าความลึกของต้นไม้จำแนกหมวดหมู่วิกิพีเดีย ตำแหน่ง และการกระจายตัว เช่น ตำแหน่งแรกที่คำนั้นปรากฏในเอกสาร ระยะห่างระหว่างตำแหน่งแรกกับตำแหน่งสุดท้าย เป็นต้น คุณลักษณะดังกล่าวข้างต้นนี้ จะพิจารณาจากแคนดิเดตคำสำคัญต่างๆ ที่พบในเอกสาร และจะถูกนำไปใช้เป็นข้อมูลฝึกสอนเพื่อสร้างตัวจำแนก ตรวจจับการเชื่อมโยง (link detection classifier) ต่อไป

สำหรับความแตกต่าง อย่างเห็นได้ชัดเจน ระหว่างงานวิจัยนี้ กับกรรมวิธีที่ถูกนำเสนอในงานวิจัยต้นแบบ [3],[4] ก็คือ เราได้ออกแบบระบบให้รวมเอาทั้งส่วนสกัดคำสำคัญ และส่วนแก้ไขปัญหาคำกำกวม เข้าพิจารณาพร้อมๆ เป็นขั้นตอนเดียวกัน โดยระบบจะสกัดคุณลักษณะของแคนดิเดตคำสำคัญก่อนเป็นลำดับแรก และถ้าแคนดิเดตคำสำคัญนั้น ปรากฏว่าเป็นคำกำกวม ระบบจะพิจารณาเลือกบทความวิกิพีเดียปลายทาง โดยพิจารณาจากค่าความคล้ายคลึงกันเชิงความหมาย (semantic relatedness) เพื่อสกัดคุณลักษณะเพิ่มเติม ระหว่างคู่ของแคนดิเดตคำสำคัญ กับบทความปลายทางที่มีความคล้ายคลึงกันเชิงความหมายนั้น เพื่อนำไปใช้เป็นข้อมูลฝึกสอนให้กับตัวจำแนกเครื่องจักรเรียนรู้ เพื่อเรียนรู้วิธีที่จะแนะนำการเชื่อมโยงในเอกสารใหม่ ต่อไป

บทที่ 3
วิธีการแนะนำการเชื่อมโยง

ขั้นตอนการทำงานของวิธีการแนะนำการเชื่อมโยงที่นำเสนอ จะประกอบไปด้วยกันทั้งหมดสามขั้นตอน ดังนี้คือ ขั้นตอนการเตรียมข้อมูลก่อนการประมวลผล ขั้นตอนการสกัดคุณลักษณะ ขั้นตอนการสร้างข้อมูลฝึกสอน และข้อมูลทดสอบเครื่องจักรเรียนรู้

3.1 การเตรียมข้อมูลก่อนการประมวลผล

ในงานวิจัยนี้จะใช้ข้อมูลจากวิกิพีเดียภาษาไทยเมื่อวันที่ 9 กรกฎาคม 2555 เป็นฐานข้อมูลหลักซึ่งชื่อบทความและเนื้อหาของวิกิพีเดียจะถูกจัดเก็บอยู่ในรูปของไฟล์เอ็กซ์เอ็มแอล (XML) ดังภาพที่ 3.1ซึ่งระบบวิกิพีเดียเองได้สำรองข้อมูลวิกิพีเดียภาษาไทยสามารถดาวน์โหลดได้ที่ http://dumps.wikipedia.org/thwiki/ ชื่อไฟล์ pages-articles.xml.bz2 จากไฟล์ข้อมูลที่ได้จะถูกนำมาผ่านขั้นตอนวิธีการเพื่อให้ได้ข้อมูลที่จำเป็นสำหรับนำมาใช้งานในระบบแนะนำการเชื่อมโยงวิกิพีเดียภาษาไทยดังนี้

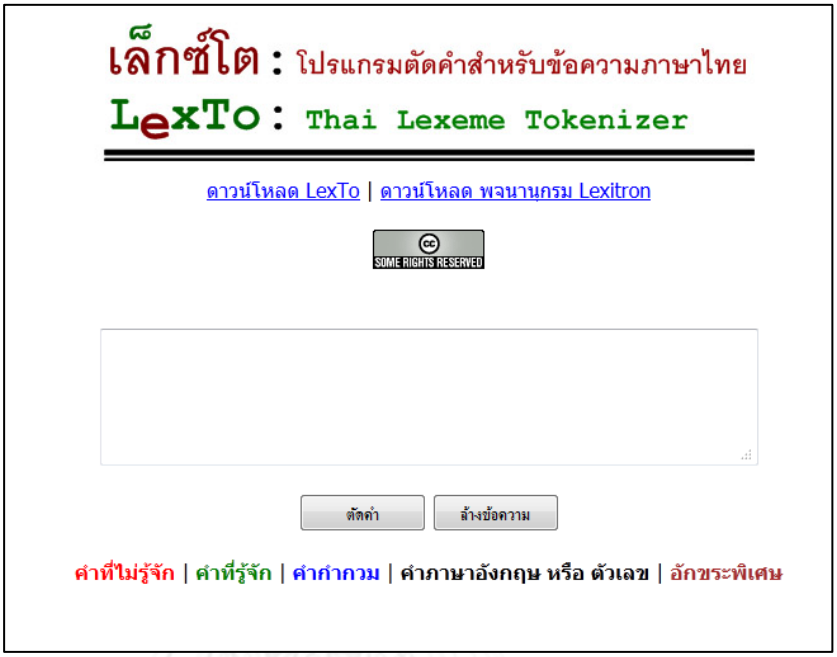
```
<page>
  <title>แคลคูลัส</title>
  <ns>0</ns>
  <id>3697</id>
  <revision>
    <id>4097607</id>
    <parentid>4084087</parentid>
    <timestamp>2012-07-03T05:12:13Z</timestamp>
    <contributor>
      <username>MerlIwBot</username>
      <id>107698</id>
    </contributor>
    <minor />
    <comment>โรบอต เพิ่ม: [[ast:Cálculu]]</comment>
    <sha1>6s1u641jlrk0r8ct872mw2nif1bks6o</sha1>
    <text xml:space="preserve">{{ต้องการอ้างอิง}}
{{แคลคูลัส}}
{{ความหมายอื่น}}
'''แคลคูลัส''' เป็นสาขาหลักของ[[คณิตศาสตร์]]ซึ่งพัฒนามาจาก[[พีชคณิต]] [[เรขาคณิต]] และปัญหาทาง[[ฟิสิกส์]] แคลคูลัสมีต้นกำเนิดจากสองแนวคิดหลัก ดังนี้

แนวคิดแรกคือ '''แคลคูลัสเชิงอนุพันธ์ (Differential Calculus)''' เป็นทฤษฎีที่ว่าด้วยอัตราการเปลี่ยนแปลง
และเกี่ยวข้องกับการหา[[อนุพันธ์]]ของ[[ฟังก์ชัน]]ทางคณิตศาสตร์ ตัวอย่างเช่น การหา [[ความเร็ว]], [[ความเร่ง]] หรือ[[ความชัน]]ของ[[เส้นโค้ง]]
บนจุดที่กำหนดให้. ทฤษฎีของอนุพันธ์หลายส่วนได้แรงบันดาลใจจากปัญหาทางฟิสิกส์
```

ภาพที่ 3.1 ตัวอย่างของบทความวิกิพีเดีย

1. การตัดคำด้วยโปรแกรมเล็กโตแบบอิงพจนานุกรม

   จากข้อเท็จจริงที่ว่า ฐานข้อมูลวิกิพีเดียภาษาไทยที่ใช้นี้ มีทั้งคำที่เป็นทั้งภาษาไทย และภาษาอังกฤษ โดยที่ภาษาไทยจะไม่มีขอบเขตของคำที่แน่ชัด ดังนั้นเราจึงเริ่มจากการตัดคำ โดยเลือกใช้วิธีการตัดคำ ทั้งภาษาไทย และภาษาอังกฤษ แบบอิงพจนานุกรมผ่านการใช้โปรแกรมเล็กโต (Lexto) [7] ดังภาพที่ 3.2 จะต้องใช้พจนานุกรมภาษาไทยและพจนานุกรมภาษาอังกฤษ โดยทีทุกคำที่พบในวิกิพีเดีย ในตำแหน่งของชื่อบทความ และข้อความแองเคอร์ จะถูกนำมาใช้สร้างพจนานุกรมคำศัพท์ควบคุม (controlled

vocabulary) ซึ่งจะถูกนำไปใช้ต่อในภายหลังกับส่วนของการสกัดคุณลักษณะของข้อมูลฝึกสอน และข้อมูลทดสอบ

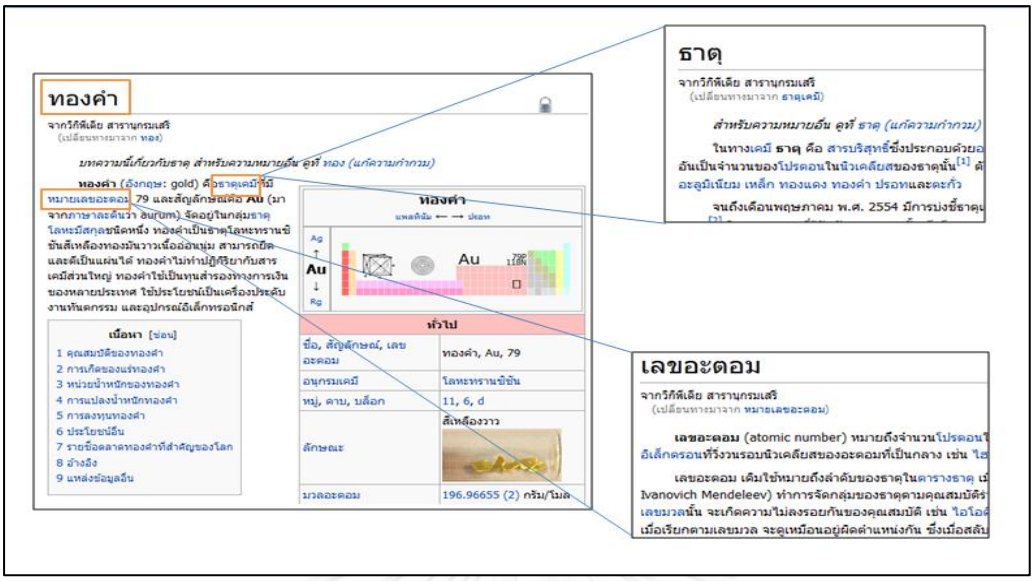


ภาพที่ 3.2 ตัวอย่างโปรแกรมเล็กโต

2. การสร้างพจนานุกรมคำศัพท์ควบคุม

พจนานุกรมคำศัพท์ (controlled vocabulary) คือ ชื่อบทความ และข้อความแองเคอร์ ดังภาพที่ 3.3 จะแสดงตัวอย่างของคำที่จะถูกนำมาใช้สร้างพจนานุกรมคำศัพท์ซึ่งพจนานุกรมนี้จะถูกนำไปใช้ร่วมกับพจนานุกรมภาษาไทยและ พจนานุกรมภาษาอังกฤษ เพื่อให้ขั้นตอนการตัดคำที่พบในเนื้อหาของบทความต่าง สามารถตัดคำได้เหมาะสม ซึ่งจะเป็นการเพิ่มประสิทธิภาพให้กับระบบแนะนำการเชื่อมโยงวิกิพีเดียภาษาไทยสามารถระบุคำได้ตรงกับคำสำคัญมากยิ่งขึ้น

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University



ภาพที่ 3.3 ตัวอย่างของคำศัพท์ควบคุม

3. การสร้างค่าสถิติที่จำเป็นสำหรับระบบแนะนำการเชื่อมโยง

ขั้นตอนของการเตรียมข้อมูลนี้ มีความจำเป็นที่จะต้องพิจารณาเก็บค่าสถิติของทุกแคนดิเดตคำสำคัญและค่าสถิติของทุกบทความซึ่งจะแสดงตัวอย่างของค่าสถิติในตาราง 3.1 ประกอบไปด้วยค่าสถิติดังต่อไปนี้

- ค่าความถี่ของคำ (Term frequency)
- ความถี่ของเอกสารที่คำนั้นปรากฏ (Document frequency)
- จำนวนครั้งที่ปรากฏของคำหรือวลีนั้นๆทั้งที่ถูกระบุเป็นคำสำคัญ (Number of to do Keyword)
- จำนวนครั้งที่ปรากฏของคำหรือวลีนั้นๆ แต่ไม่ระบุเป็นคำสำคัญ (Number of not to do Keyword)
- คำสำคัญ กำกวม หรือไม่ (Keyword is Ambiguous)
- คำสำคัญเขียนด้วยภาษาอังกฤษ หรือไม่ (Keyword is English word)
- คำสำคัญเขียนด้วยภาษาไทย หรือไม่ (Keyword is Thai word)
- คำสำคัญพบที่ชื่อบทความ หรือไม่ (Keyword is Article name)
- คำสำคัญพบที่แองเคอร์ หรือไม่ (Keyword is Anchor Text)
- รายการของบทความวิกิพีเดียปลายทางที่คำสำคัญนั้นเชื่อมโยง (List of Destination page of keyword)
- รายการของบทความวิกิพีเดียที่ปรากฏคำสำคัญนั้นๆ (List of Source page of keyword)
- ความยาวของบทความ (จำนวนคำ) (Length of Source page)
- รายการของบทความวิกิพีเดียที่ชี้เข้าหาบทความนั้นๆ (List of In link)

- รายการของบทความวิกิพีเดียที่บทความนั้นๆชี้ออก (List of Out link)
- บทความวิกิพีเดียนั้นๆ เป็นหน้าแก้ไขคำกำกวม หรือไม่ (Articles is Disambiguation page)
- บทความวิกิพีเดียนั้นๆ เป็นหน้าเปลี่ยนทาง หรือไม่ (Articles is Redirect page

ตารางที่ 3.1 ตัวอย่างของค่าสถิติของคำและบทความ

คำหรือวลี	ค่าสถิติของคำ	ค่าสถิติของบทความ
รัฐวอชิงตัน	Term frequency = 75 Document frequency = 63 Count to do Keyword = 53 Ambiguous Word = false English Word = false Thai Word = true Article Name = true List of Source page of keyword = [244360, 823, 244361, 119117, 4697, 194538, 56151, 114430, 34529, 205696, 315495, 258376, 4070, 73220, 10820, 73037, 180195, 19108, 310934, 103205, 132126, 10336, 276337, 235948, 14515, 142199, 98996, 187063, 202189, 214038, 63019, 85789, 191376, 283828, 219264, 114841] List of Destination page of keyword = (7866=53)	Article ID = 7866 Length of Source Page = 4,788 Disambiguation Page = false Redirect Page = false Out link = 19 List of In link = [244360, 823, 244361, 119117, 4697, 194538, 56151, 114430, 34529, 205696, 315495, 258376, 4070, 73220, 10820, 73037, 19108, 180195, 310934, 103205, 132126, 10336, 235948, 14515, 276337, 142199, 98996, 198799, 187063, 124597, 202189, 122376, 214038, 63019, 7613, 85789, 191376, 283828, 219264, 114841]
ประเทศไทย	Term frequency = 10,434 Document frequency = 6,683 Count to do Keyword = 3,888 Ambiguous Word = false English Word = false Thai Word = true Article Name = true List of Source page of keyword =	Article ID = 936 Length of Source Page = 63,544 Disambiguation Page = false Redirect Page = false List of Out link = List of In link = [195718, 10999, 8226, 132519, 323200, 47214, 291147, 6866, 310751, 195711, 319702, 320291,

	[195718, 10999, 8226, 132519, 323200, 47214, 291147, 310751, 195711, 319702, 320291, 181939, 6882, 275950, 291922, 6097, 6879, 63929, 113459, 102216, 6081, 102253, 238100, 187605, 51498, 98539, 5353, 287068, 196013, 52828, 321789, 1847, 196276, 148586, 64593, 168007, 8231, 169549, 140582, 136650, 8230, 63019, 63904, 40399, 310747, 308851, 1836, 10950, 10069, 63859, 5331, 235396, 313013, 306627, 7592, 167310, 86086, 100016, 185258, 58868, 205203, 109302, 100013, 310799, 177390, 40388, 2101, 251327, 181196, 264351, 2106, 33594, 35700, 165739, 35709, 156458, 100032, 63970, 295237, 63037, 218716, 169565, 100027, 147912, 101157, 321774, 258435, 62309, 328212, 323223, 135044, 312893, 63041, 101131, 196230, 313937, 196237, 291983, 291985, 47225, 323210, 156479, 181998, 242387, 122322, ..] List of Destination page of keyword = 121526=1, 109655=1, 147002=4, 99304=6, 53821=2, 936=3874	181939, 6882, 275950, 291922, 6097, 6879, 63929, 120167, 113459, 102253, 187605, 51498, ..]
singleplayer	Term frequency = 39 Document frequency = 39 Count to do Keyword = 37 Ambiguous Word = false English Word = true Thai Word = false	Article ID = 204211 Length of Source Page = 39 Disambiguation Page = false Redirect Page = true [204208] Out link = 1 List of In link = [255306, 253613, 262831,

	Article Name = Single-player List of Source page of keyword = [255409, 200023, 178626, 177889, 255139, 227543, 195300, 177886, 227799, 228772, 200218, 229134, 200219, 183566, 179911, 198588, 230627, 255223, 200532, 183223, 179961, 255295, 310443, 177908, 178513, 178960, 262335, 185809, 259789, 264905, 186576, 181852, 227549, 312445, 252644, 185412, 255168] List of Destination page of keyword = 204211=1, 262489=36	291750, 228942, 64626, 287826, 222068, 178960, 19990, 216410, 262601, 259691, 288767, 255641, 20760, 207438, 20672]

4. การสร้างตัวอย่างบวกและตัวอย่างลบ

การที่จะให้เครื่องจักรเรียนรู้สามารถทำนายว่าคำที่ปรากฏในบทความควรที่ถูกแนะนำให้สร้างการเชื่อมโยงนั้นจะมีทั้งข้อมูลที่เป็นทั้งตัวอย่างบวกและตัวอย่างลบเพื่อให้ระบบเรียนว่าข้อมูลลักษณะใดควรที่จะถูกแนะนำการเชื่อมและข้อมูลลักษณะใดไม่ควรที่จะถูกแนะนำการเชื่อมโยง ซึ่งข้อมูลในวิกิพีเดียภาษาไทยนั้นมีลักษณของข้อมูลทั้งในแบบของตัวอย่างบวกและตัวอย่างลบ

- ตัวอย่างบวก (Positive example) คือ ตัวอย่างของลักษณะที่มีสร้างการเชื่อมโยงกันระหว่างคำสำคัญกับบทความปลายทางจากลักษณข้อมูลดังกล่าว เราจึงนำการเชื่อมโยงดังกล่าวมาสร้างเป็นข้อมูลตัวอย่างบวกเพื่อใช้ในการฝึกสอนให้ระบบทำนายว่าคำลักษณะดังกล่าวระบบควรแนะนำการเชื่อมโยง โดยที่ระบบจะเก็บข้อมูลตัวอย่างบวกในลักษณะดังตารางที่ 3.2

ตารางที่ 3.2 ตัวอย่างของลักษณะข้อมูลที่เป็นตัวอย่างบวก

คำหรือวลี	**บทความต้นทาง**	**บทความปลายทาง**
รัฐ	รัฐวอชิงตัน	รัฐ
สหรัฐอเมริกา	รัฐวอชิงตัน	สหรัฐอเมริกา
มหาสมุทรแปซิฟิก	รัฐวอชิงตัน	มหาสมุทรแปซิฟิก

โบอิง	รัฐวอชิงตัน	โบอิง
รัฐอิลลินอยส์	รัฐวอชิงตัน	รัฐอิลลินอยส์
ไมโครซอฟท์	รัฐวอชิงตัน	ไมโครซอฟท์
แอมะซอนคอม	รัฐวอชิงตัน	แอมะซอนคอม
นินเทนโดอเมริกา	รัฐวอชิงตัน	นินเทนโดอเมริกา

- ตัวอย่างลบ (Negative example) คือ ตัวอย่างของลักษณะที่ไม่ได้สร้างการเชื่อมโยงกันระหว่างคำสำคัญกับบทความปลายทาง ในงานวิจัยนี้จะนำข้อมูลดังกล่าวมาสร้างเป็นข้อมูลตัวอย่างลบเพื่อใช้ในการฝึกสอนให้ระบบทำนายว่าคำลักษณะดังกล่าวระบบไม่ควรแนะนำการเชื่อมโยงโดยที่ระบบจะเก็บข้อมูลตัวอย่างลบในลักษณะดังตารางที่ 3.3

ตารางที่ 3.3 ตัวอย่างของลักษณะข้อมูลที่เป็นตัวอย่างลบ

คำหรือวลี	**บทความต้นทาง**	**บทความปลายทาง**
เมืองซีแอตเติล	รัฐวอชิงตัน	Null
โลก	รัฐวอชิงตัน	Null
ประเทศ	รัฐวอชิงตัน	Null
ข้อมูล	รัฐวอชิงตัน	Null
กล่อง	รัฐวอชิงตัน	Null
จำนวนประชากร	รัฐวอชิงตัน	Null

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จากขั้นตอนการเตรียมข้อมูลที่กล่าวมาข้างต้นนั้นเราจะต้องแปลงข้อมูลวิกิพีเดียที่อยู่ในรูปของไฟล์เอ็กซ์เอ็มแอล (XML) ด้วยการสร้างพจนานุกรมคำศัพท์ควบคุมแล้วจึงกำหนดให้คำที่ปรากฏในพจนานุกรมคำศัพท์ควบคุมทั้งหมด เป็นแคนดิเดตคำสำคัญ ดังนั้นในระหว่างการวิเคราะห์เอกสารวิกิพีเดียทั้งหมดในขั้นตอนของการเตรียมข้อมูลนี้ เราจะพิจารณาเก็บค่าสถิติของทุกแคนดิเดตคำสำคัญ และสร้างตัวอย่างบวกและตัวอย่างลบจากข้อมูลการเชื่อมโยงกันของระว่างบทความต้นทางและบทความปลายทางผ่านคำสำคัญที่ จากข้อมูลที่เตรียมในขั้นตอนที่ 3.1 จะถูกจัดเก็บเพื่อนำไปใช้ประโยชน์ในส่วนถัดไป โดยที่ขั้นตอนถัดไป คือ การอธิบายการสกัดคุณลักษณะที่จะใช้สำหรับฝึกสอนเครื่องจักรเรียนรู้

3.2 การสกัดคุณลักษณะ

เมื่อเราสร้างพจนานุกรมคำศัพท์ควบคุมเรียบร้อยแล้ว ทุกคำหรือวลีในเอกสารที่ถูกพบว่าตรงกับคำศัพท์ควบคุม จะถูกเลือกให้เป็นแคนดิเดตคำสำคัญทั้งหมด เนื่องจากวิธีการแนะนำการเชื่อมโยงของเราได้ใช้ตัวจำแนกเครื่องจักรเรียนรู้ในการเลือกคำในเอกสารเพื่อแนะนำการเชื่อมโยง ดังนั้นเราจึงต้องสกัดคุณลักษณะเวกเตอร์ (feature vector) ของทุกๆ แคนดิเดตคำสำคัญเหล่านั้น สำหรับทั้งข้อมูลฝึกสอน และข้อมูลทดสอบ ในที่นี้ เราจะขออธิบายแต่ละคุณลักษณะ ดังนี้

3.2.1 ความน่าจะเป็นของการทำลิงก์

เราได้นำแนวคิดการคำนวณค่าความน่าจะเป็นของการทำลิงค์จากงานวิจัย [3] มาใช้ด้วย ค่าความน่าจะเป็นของการทำลิงค์ของแต่ละแคนดิเดต W ที่ถูกเลือกให้เป็นคำสำคัญในเอกสาร นิยามได้ว่า เป็นจำนวนของเอกสารที่คำนั้นปรากฏ และถูกทำเป็นลิงก์ ($count(D_{key})$) หารด้วยจำนวนเอกสารทั้งหมดที่คำนั้นปรากฏ ($count(D_W)$) ค่าสถิตินี้สามารถคำนวณได้จากบทความวิกิพีเดีย

$$\boldsymbol{LP(Keyword|W) \approx \frac{count(D_{key})}{count(D_W)}} \quad (2)$$

3.2.2 ความคล้ายคลึงเชิงความหมาย

สำหรับคำที่มีความกำกวม การที่จะเลือกบทความปลายทางให้เหมาะสมนั้นสามารถเลือกได้จากค่าความคล้ายคลึงกันเชิงความหมาย ระหว่างเนื้อหาบทความปลายทางกับเนื้อหารอบๆ คำกำกวมนั้นๆ ซึ่งการคำนวณความคล้ายคลึงกันระหว่างบทความ *A* และ *B* ในที่นี้เราใช้วิธีการดั้งเดิมที่ใช้ในงานวิจัยด้านการสืบค้นข้อมูล โดยวัดความคล้ายคลึงกันระหว่างเอกสารแบบโคไซต์ [13] ดังนี้

$$\boldsymbol{Similarity(A,B) = \frac{A \cdot B}{\|A\|\|B\|}} \quad (3)$$

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

โดยในที่นี้ *M* และ *N* คือสองเอกสารที่กำลังสนใจ ซึ่งจะถูกแสดงโดยรูปของเวกเตอร์ของคำและมีค่าแต่ละมิติของเวกเตอร์ เป็นค่าน้ำหนักแบบ *tf.idf* [12]

ส่วนการพิจารณาเลือกเอกสารเพื่อการแก้ไขปัญหาคำกำกวม เราพิจารณาจากคำที่ไม่กำกวมรอบๆ คำกำกวมนั้น จำนวน 3 คำ (หน้าและหลัง) ในการที่จะเลือกเอกสารที่มีความหมายที่เหมาะสมที่สุดให้กับคำกำกวมนั้น เราจะพิจารณาจากค่าความน่าจะเป็นของการเชื่อมโยงไปยังบทความปลายทาง กับค่าความคล้ายคลึงกันของเอกสารต้นทางกับปลายทาง ตามสมการ (3)

$$\boldsymbol{Destination(C,k) = \underset{b_j \in B}{argmax}[P(k,b_j) \cdot sim(C,b_j)]} \quad (4)$$

โดยในที่นี้ *C* คือเอกสารนำเข้าต้นทาง *k* คือ คำสำคัญที่กำลังพิจารณา *B* คือชุดของบทความวิกิพีเดียที่เป็นไปได้ทั้งหมดของ *k* ส่วน $P(k,b_j)$ คือ ค่าความน่าจะเป็นของการเชื่อมโยงไปยังบทความปลายทาง b_j สำหรับคำสำคัญ *k*

3.2.3 ตำแหน่งที่ปรากฏ

เพื่อหลีกเลี่ยงปัญหาการแนะนำการเชื่อมโยงมากเกินความจำเป็น การทำลิงค์ไปยังบทความอื่นๆ ในวิกิพีเดียควรเลือกเฉพาะตำแหน่งที่ต้องการชี้แจงส่วนเนื้อหาที่จำเป็นเพิ่มเติม ที่ถูกพบเป็นครั้งแรกเท่านั้น ในที่นี้เราให้เครื่องจักรเรียนรู้ ได้เรียนรู้ว่า ตำแหน่งใดบนเอกสารสมควรที่จะถูกเลือกทำลิงค์ ตัวอย่างเช่น การพบคำสำคัญใดๆ ครั้งแรก จะถูกระบุค่าตำแหน่งเป็น 1 และจะถูกระบุค่าเป็น 2 และ 3 เมื่อคำสำคัญคำเดิมนั้นถูกพบอีกเป็นครั้งที่สอง และสาม ในเอกสารที่กำลังพิจารณาตามลำดับ

3.2.4 คุณลักษณะอื่นๆ

คุณลักษณะอื่นๆ ที่จำเป็น ได้แก่ คุณลักษณะที่ระบุว่าเป็นชื่อบทความวิกิพีเดียหรือไม่ เป็นคำสำคัญที่เขียนด้วยภาษาอังกฤษหรือไม่ เป็นคำกำกวมหรือไม่ เป็นหน้าเปลี่ยนทางหรือไม่ เป็นต้น

3.3 การสร้างข้อมูลฝึกสอน และทดสอบ

ทั้งในขั้นตอนการฝึกสอนตัวจำแนกด้วยเครื่องจักรเรียนรู้ เราจะสกัดคุณลักษณะจากคู่ระหว่างแคนดิเดตคำสำคัญ กับบทความวิกิพีเดียปลายทางที่เหมาะสมจากตัวอย่างเอกสารวิกิพีเดียฝึกสอนทั้งหมด คุณลักษณะที่สกัดได้จะแสดงอยู่ในรูปเวกเตอร์ตามตัวอย่างแสดงดังภาพที่ 3.4 ในรูปเออาร์เอฟเอฟ (Attribute-Relation File Format) [6]

ภาพที่ 3.4 ตัวอย่างของข้อมูลคุณลักษณะ

```
1:      @attribute link_prob numeric
2:      @attribute relatedness numeric
3:      @attribute position numeric
4:      @attribute is_title {true,false}
5:      @attribute is_english {true,false}
6:      @attribute is_ambiguous {true,false}
7:      @attribute is_redirect {true,false}
8:      @atribute 'link' {Y,N}
9:      @data
10:     0.154639,0.121461,1,FALSE,FALSE,TRUE,FALSE,Yes
11:     0.214286,0.142751,1 FALSE,FALSE,TRUE,FALSE,Yes
12:     0.194245,0.090028,1 FALSE,FALSE,TRUE,FALSE,Yes
13:     0.5,0.089857,1,FALSE,FALSE,TRUE,FALSE,Yes
14:     0.563218,0.071167,1,FALSE,FALSE,TRUE,FALSE,Yes
15:     0.586207,0.367333,1,FALSE,FALSE,TRUE,FALSE,Yes
16:     0.028523,0.600061,1,FALSE,FALSE,TRUE,FALSE,Yes
```

จากรูปที่ 1 อธิบายเพิ่มเติมเวคเตอร์คุณลักษณะได้ว่า บรรทัดที่ 1 แสดงค่าความน่าจะเป็นที่ถูกทำลิงค์ บรรทัดที่ 2 แสดงค่าความคล้ายคลึงเชิงความหมาย (คำนวณโดยสมการที่ 2) ระหว่างบทความปลายทาง กับแคนดิเดตคำสำคัญที่กำลังพิจารณา ส่วนบรรทัดอื่นแสดงตัวอย่างของคุณลักษณะอื่นๆ ที่อธิบายในหัวข้อ 3.2 ที่ผ่านมา

ในขั้นตอนของการทดสอบก็เช่นกัน คุณลักษณะที่สกัดได้จากคู่ระหว่างแคนดิเดตคำสำคัญ กับบทความวิกิพีเดียปลายทางที่เหมาะสม จะถูกส่งไปให้ตัวจำแนกทำการตัดสินใจว่า การเชื่อมโยงดังกล่าวสมควรที่จะถูกแนะนำหรือไม่

3.4 การทดสอบประสิทธิผลของระบบ

การวัดผลประสิทธิภาพของระบบแนะนำการเชื่อมโยงวิกิพีเดียภาษาไทย (Link suggestion) จะแบ่งออกเป็น 2 ส่วน ส่วนแรกจะทดสอบในส่วนของการแนะนำคำสำคัญ ส่วนที่สองจะทดสอบความถูกต้องโดยรวมของระบบแนะนำการเชื่อมโยง

3.4.1 การทดสอบประสิทธิภาพของระบบแนะนำการเชื่อมโยง

จะวัดประสิทธิภาพของระบบว่าคำสำคัญควรแนะนำการเชื่อมโยงหรือไม่ การทดสอบจะแบ่งออกเป็น 3 รูปแบบ ได้แก่

1. True Positive (TP) คือ ผลลัพท์ที่ระบบเชื่อมโยงคำสำคัญกับบทความปลายทางได้ถูกต้องเมื่อเทียบกับตัวอย่าง
2. False Positive (FP) คือ ผลลัพท์ที่ระบบเชื่อมโยงคำสำคัญกับบทความปลายทางไม่ถูกต้องเมื่อเทียบกับตัวอย่าง
3. False Negative (FN) คือ ผลลัพท์ที่ระบบไม่ได้เชื่อมโยงคำสำคัญกับบทความปลายทาง และไม่ถูกต้องเมื่อเทียบกับตัวอย่าง

ผลลัพท์ทั้ง 3 รูปแบบที่ได้จากการทดสอบ เราจะนำมาคำนวนผ่านสมการ ดังนี้

1. ค่าความแม่นยำ (Precision) คือ ค่าที่บ่งบอกว่าระบบมีความแม่นยำในการแนะนำการเชื่อมโยงจากกลุ่มคำตอบได้มากน้อยเพียงใด

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

2. ค่าระลึก (Recall) คือ ค่าที่บ่งบอกว่าระบบแนะนำการเชื่อมโยงได้ครอบคลุมคำตอบทั้งหมดมากน้อยเพียงใด

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

3. ค่าความถูกต้อง (F-measure) คือ สมการที่นำค่าความแม่นยำและค่าระลึกมาคำนวนเพื่อระบุถึงประสิทธิภาพโดยรวมของระบบ

$$\boldsymbol{F - measure} = \frac{\boldsymbol{2 \cdot Precision \cdot Recall}}{\boldsymbol{Precision + Recall}} \quad (7)$$

3.4.2 การทดสอบประสิทธิผลของระบบแนะนำการเชื่อมโยง

จะวัดประสิทธิภาพของระบบว่าระบบแนะนำการเชื่อมโยงได้ถูกต้องมากน้อยเพียงใดเมื่อเทียบกับตัวอย่าง ซึ่งจะวัดจากค่าความถูกต้อง (Accuracy)

$$\boldsymbol{Accuracy} = \frac{\text{จำนวนลิงก์ที่เชื่อมโยงบทความปลายทางได้ถูกต้องโดยระบบ}}{\boldsymbol{TP + FP}} \quad (8)$$

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 4
การพัฒนาเครื่องมือสนับสนุนระบบการแนะนำการเชื่อมโยง

เพื่อทำให้งานวิจัยมีความสมบูรณ์มากขึ้น งานวิจัยนี้จึงได้จัดทำเครื่องมือเพื่อสนับสนุนวิธีการแนะนำการเชื่อมโยงวิกิพีเดียภาษาไทย โดยทำการพัฒนาด้วยภาษาจาวา (Java Programming) ร่วมกับการใช้งานไลบรารี (Library) ที่มีการพัฒนาไว้ใช้งานอยู่แล้วมาประยุกต์เพื่อให้การทำงานของเครื่องมือทำงานได้อย่างมีประสิทธิภาพมากขึ้น โดยขั้นตอนในการพัฒนาเครื่องมือมีดังต่อไปนี้

4.1 สภาพแวดล้อมและเครื่องมือที่ใช้ในการพัฒนา

สภาพแวดล้อมทางด้านฮาร์ดแวร์และซอฟต์แวร์ที่ใช้ในการพัฒนาระบบมีดังต่อไปนี้

1. ฮาร์ดแวร์
   - หน่วยประมวลผล อินเทล 2.50 กิกะเฮิร์ตซ์ (Core i5-2520 2.50 GHz.)
   - หน่วยความจำ (RAM) 8.0 กิกะไบต์ (7.89 GB)
   - ฮาร์ดดิสก์ (Hard disk) 500 กิกะไบต์ (500 GB)
2. ซอฟต์แวร์
   - ระบบปฏิบัติการ วินโดวส์เซเวน โฮม(Windows 7 Home)
   - เครื่องมือพัฒนาโปรแกรมภาษาวาจา ประกอบด้วยรายการต่าง ๆ ดังนี้
     1. โปรแกรมอิคลิปส์เวอร์ชันเฮลิออส (Eclipse Java EE IDE for Web Developers, Version: INDIGO Release)
     2. จาวาเอพีไอไลบรารี (Java API Library) ประกอบด้วย
        - จาวาซิสเต็มไลบรารี เวอร์ชัน 1.6 (Java JRE System Library Version 1.6)
3. การติดตั้งซอฟต์แวร์ในการพัฒนาระบบ

   เมื่อได้กำหนดเครื่องมือสำหรับการพัฒนาระบบเรียบร้อยแล้ว ขั้นต่อไปคือการติดตั้งเครื่องมือทั้งหมดลงในเครื่องคอมพิวเตอร์ที่ใช้พัฒนาระบบ โดยมีลำดับการติดตั้งเครื่องมือเป็นไปตามขั้นตอนต่อไปนี้
   - ติดตั้งระบบปฏิบัติการ วินโดวส์เซเวน โฮม
   - ติดตั้งชุดพัฒนาโปรแกรมภาษาวาจา เวอร์ชัน 1.6

บทที่ 5
การทดสอบ

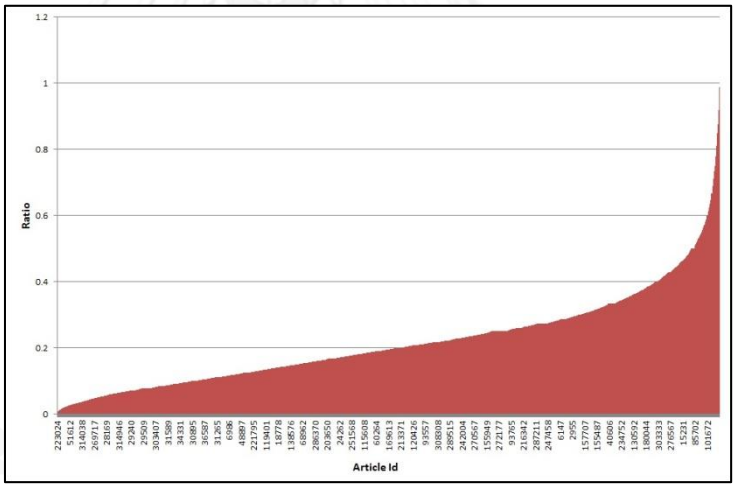
เมื่อทำการพัฒนาเครื่องมือเรียบร้อยแล้วจะต้องทำการทดสอบการทำงานของเครื่องมือว่าเครื่องมือนั้นสามารถทำงานได้อย่างถูกต้องตามแบบจำลองที่กำหนดไว้และ เราจะต้องทำการเตรียมข้อมูลจากวิกิพีเดียภาษาไทย เดือนกรกฎาคม 2555 ตามกรรมวิธีที่ได้อธิบายไว้ในหัวข้อ 3.1 ที่ผ่านมา พจนานุกรมคำศัพท์ควบคุมที่สังเคราะห์ได้จากชื่อบทความ และข้อความแองเคอร์ของวิกิพีเดีย มีจำนวนคำศัพท์เป็นจำนวนมากกว่าหนึ่งแสนคำ ครบถ้วนตามความต้องการด้านหน้าที่หรือไม่

5.1 ข้อมูลที่ใช้ฝึกสอน และทดสอบระบบ

5.1.1. การเลือกบทความเพื่อใช้ในการสร้างชุดข้อมูลฝึกสอนและทดสอบ

จากข้อมูลวิกิพีเดียภาษาไทย เดือนกรกกฎาคม 2555 เราสกัดบทความทั้งหมดได้ 132,385 บทความ จากบทความทั้งหมดเราใช้วิธีการเลือกโดยสนใจที่จำนวนลิงก์ชี้ออก และจำนวนคำสำคัญของแต่ละบทความหรือเอกสาร

1. ผลการคำนวนสัดส่วนระหว่างจำนวนลิงก์ขี้ออกต่อจำนวนคำสำคัญออกของบทความวิกิพีเดียภาษาไทย



ภาพที่ 5.1 แสดงภาพผลการคำนวนสัดส่วนระหว่างจำนวนลิงก์ขี้ออกต่อจำนวนคำสำคัญบทความวิกิพีเดียภาษาไทย

5.1.2. การทดสอบประสิทธิภาพของระบบแนะนำการเชื่อมโยง

การที่จะให้ตัวจำแนกเรียนรู้วิธีแนะนำการเชื่อมโยงทั้งส่วนการฝึกสอน และส่วนของการทดสอบเบื้องต้นนั้น เราจะเลือกบทความวิกิพีเดียจากขั้นตอน 5.1.1 ทำการสุ่มเลือกจำนวน 1000 บทความในแต่ละกลุ่ม เพื่อนำมาใช้เป็นข้อมูลฝึกสอนและทดสอบตัวจำแนกในส่วนสกัดคำสำคัญ และสุ่มเลือกอีกจำนวน 100 บทความในแต่ละกลุ่ม เพื่อนำมาใช้ทดสอบในส่วนแก้ไขปัญหาคำกำกวมซึ่งเราจะสนใจเพียงค่าความถูกต้อง ตัวอย่างของข้อมูลฝึกสอนแสดงอยู่ในรูปคู่ระหว่างข้อความแองเคอร์กับบทความปลายทาง สำหรับคำกำกวมนั้น คู่ระหว่างข้อความแองเคอร์กับบทความปลายทางที่

เหมาะสมจะถูกกำหนดให้เป็นตัวอย่างบวก ส่วนบทความที่เหลือที่เป็นไปได้จะถูกกำหนดให้เป็นตัวอย่างลบ และเนื่องจากโดยปกติแล้ว ผู้เขียนวิกิพีเดียจะสร้างคำอธิบายการเชื่อมโยง (link annotation) เพียงครั้งเดียวต่อหนึ่งคำสำคัญในบทความวิกิพีเดียนั้น ดังนั้นคำสำคัญเดียวกัน ถ้าปรากฏในตำแหน่งอื่นๆ จะถูกกำหนดให้เป็นตัวอย่างลบด้วยเช่นเดียวกัน

5.1.3. การทดสอบของระบบแนะนำการเชื่อมโยง

จากการขั้นตอนการเลือกบทความที่กล่าวในหัวข้อ 5.1.1 เราจะทำการสุ่มเลือกบทความดังกล่าวนำมาใช้เป็นข้อมูลฝึกสอนและทดสอบระบบ แบ่งบทความออกเป็นทั้งหมด 5 กลุ่ม โดยแบ่งจากค่าคะแนนสัดส่วนลิงก์ขี้ออกต่อจำนวนคำสำคัญของแต่ละบทความ กลุ่มที่ 1 เป็นบทความที่มีคะแนนสัดส่วนลิงก์ขี้ออกต่อจำนวนคำสำคัญ [0.01 – 0.04], กลุ่มที่ 2 เป็นบทความที่มีคะแนนสัดส่วนลิงก์ขี้ออกต่อจำนวนคำสำคัญ (0.041 – 0.08], กลุ่มที่ 3 เป็นบทความที่มีคะแนนสัดส่วนลิงก์ขี้ออกต่อจำนวนคำสำคัญ (0.081 – 0.12], กลุ่มที่ 4 เป็นบทความที่มีคะแนนสัดส่วนลิงก์ขี้ออกต่อจำนวนคำสำคัญ (0.121 – 0.16], และกลุ่มสุดท้ายกลุ่มที่ 5เป็นบทความที่มีคะแนนสัดส่วนลิงก์ขี้ออกต่อจำนวนคำสำคัญ (0.161 – 0.4],เราจะสุ่มเลือกบทความจากแต่ละกลุ่มมาสร้างเป็นข้อมูลฝึกสอนและทดสอบแต่ละชุดมีจำนวนบทความจำนวน 1000 บทความ

5.1.4. การทดสอบของระบบวิกิไมเนอร์

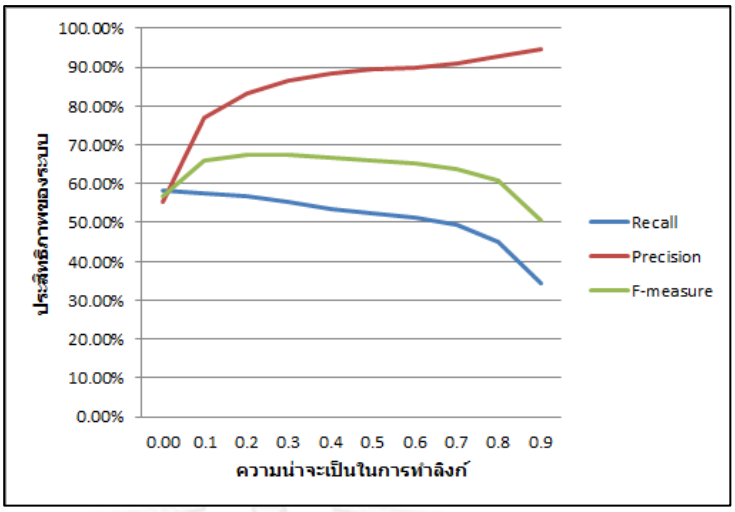
ในงานวิจัยนี้เราจะเรายังได้ทำการทดสอบข้อมูลวิกิพีเดียกับระบบ ของวิกิไมเนอร์จากบทความที่ได้เลือกไว้ในขั้นตอน 5.1.1 เราจากการขั้นตอนการเลือกบทความที่กล่าวในหัวข้อ 5.1.3 เราจะทำการสุ่มเลือกบทความดังกล่าวนำมาใช้เป็นข้อมูลฝึกสอนและทดสอบจำนวน 1000 บทความ โดยทำการทดสอบทั้งในส่วนของการสกัดคำสำคัญและการแก้ไขปัญหาความกำกวม

5.2 ผลการทดลองที่ได้

5.2.1 ผลการทดสอบระบบวิกิไมเนอร์

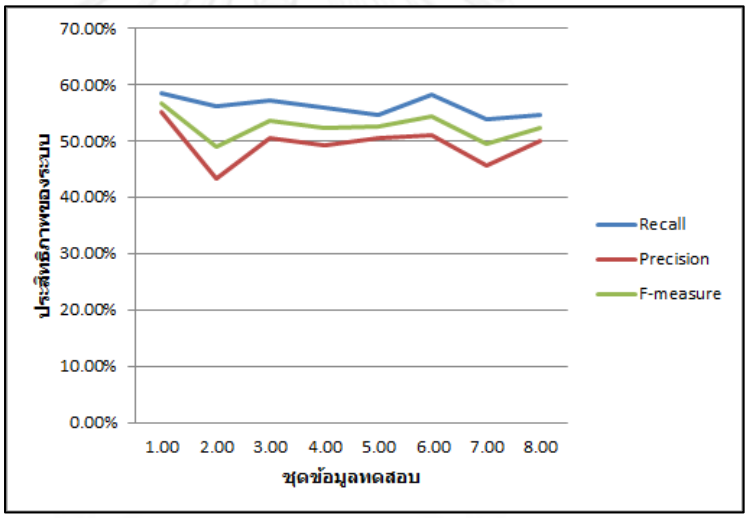
ในงานวิจัยนี้เราได้ทำการวิเคราะห์ผลที่ได้จากการทดลองบทความวิกิพีเดียภาษาไทยผ่านระบบของ วิกิไมเนอร์ [4] โดยใช้ข้อมูลชุดฝึกสอนและทดสอบจากหัวข้อ 5.1.5 และใช้ตัวจำแนกเครื่องจักรเรียนรู้แบบต้นไม้ตัดสินใจ C4.5 [14] ในโปรแกรมสำหรับสร้างเครื่องจักรเรียนรู้เวกา (WEKA machine learning toolkit) [6] โดยที่เราจะแสดงผลการทดสอบออกเป็น 2 ส่วนด้วยกัน

1. ผลการทดลองจากการเปลี่ยนค่าความน่าจะเป็นในการทำลิงก์ ดังภาพที่ 5.1



ภาพที่ 5.2 แสดงประสิทธิภาพของระบบวิกิไมเนอร์กับข้อมูลวิกิพีเดียภาษาไทย

2. ผลการทดลองจากการเปลี่ยนชุดข้อมูลทดสอบที่แตกต่างกัน ดังภาพที่ 5.2



ภาพที่ 5.3 แสดงประสิทธิภาพของระบบวิกิไมเนอร์กับข้อมูลชุดทดสอบที่แตกต่างกัน

จากผลการทดลองทั้งสอง ส่วนนี้แสดงให้เห็นว่าระบบ แนะนำการเชื่อมโยงของ วิกิไมเนอร์นั้นสามารถได้ค่าความถูกต้อง (F-measure) เฉลี่ย อยู่ที่ 50 – 60 %

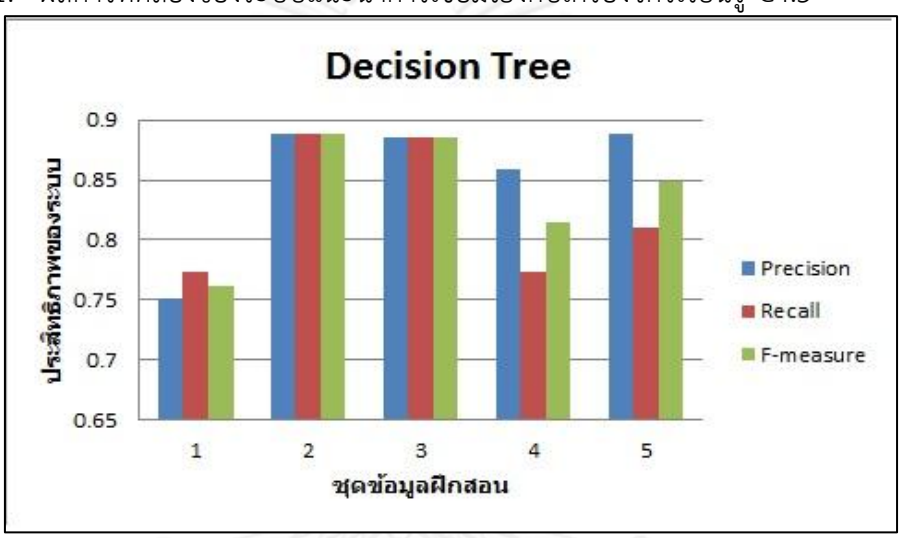
5.2.2 ผลการทดสอบระบบแนะนำการเชื่อมโยง

ถึงแม้ว่า ระบบการแนะนำการเชื่อมโยงในงานวิจัยนี้ เราจะรวมขั้นตอนการทำงาน ทั้งส่วนระบุคำสำคัญ และส่วนแก้ไขปัญหาคำกำกวม เข้าไว้เป็นขั้นตอนเดียวกันก็ตาม เพื่อให้ง่ายต่อการเปรียบเทียบผลลัพธ์ที่ได้กับงานวิจัยก่อนหน้า เราจะตรวจสอบ และรายงานการประเมินผลแยกกันเป็นสองส่วน สำหรับส่วนแรกจะใช้ ค่าความแม่นยำ (precision) และ ค่าระลึก (recall) จากจำนวนแคนดิเดตคำสำคัญที่เลือกได้อย่างถูกต้อง สำหรับในส่วนที่สอง เราจะวัดค่าความถูกต้อง (accuracy)

คือจำนวน บทความปลายทางที่เลือกได้ถูกต้องเมื่อเทียบกับจำนวนการเชื่อมโยงทั้งหมดที่ระบบแนะนำ

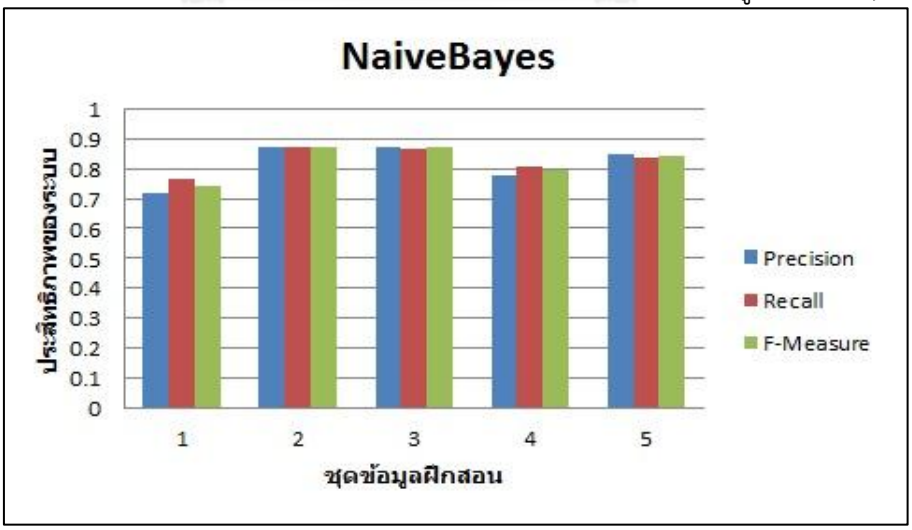
เราได้ทำการทดสอบวิธีการที่นำเสนอ โดยใช้ตัวจำแนกเครื่องจักรเรียนรู้ C4.5, Naivebayes และ Support Vector Machineสำหรับแต่ละบทความฝึกสอน และบทความทดสอบที่ได้รับการสุ่มเลือกนั้น เราจะทำการเลือกใหม่ และทำการทดลองซ้ำๆ กัน และจะเฉลี่ยผลลัพธ์ที่ได้ทั้งหมด สุดท้ายแล้ว เราได้รับผลลัพธ์ค่าความแม่นยำประมาณ 80% ค่าความระลึกประมาณ 80% และค่าความถูกต้องของระบบแนะนำการเชื่อมโยงถึง 85% คำนวณโดยเฉลี่ย

1. ผลการทดลองของระบบแนะนำการเชื่อมโยงกับเครื่องจักรเรียนรู้ C4.5



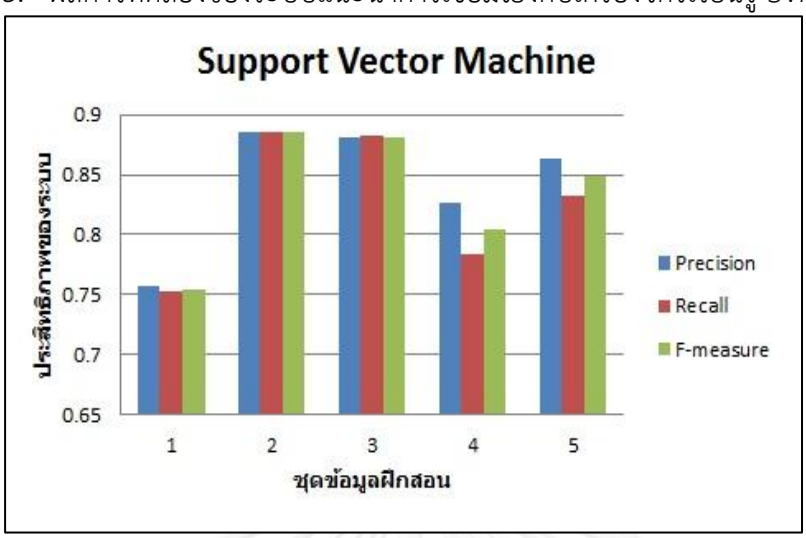
ภาพที่ 5.4 แสดงภาพผลการทดลองของระบบแนะนำการเชื่อมโยงกับเครื่องจักรเรียนรู้ C4.5

2. ผลการทดลองของระบบแนะนำการเชื่อมโยงกับเครื่องจักรเรียนรู้ Naïve Bayes



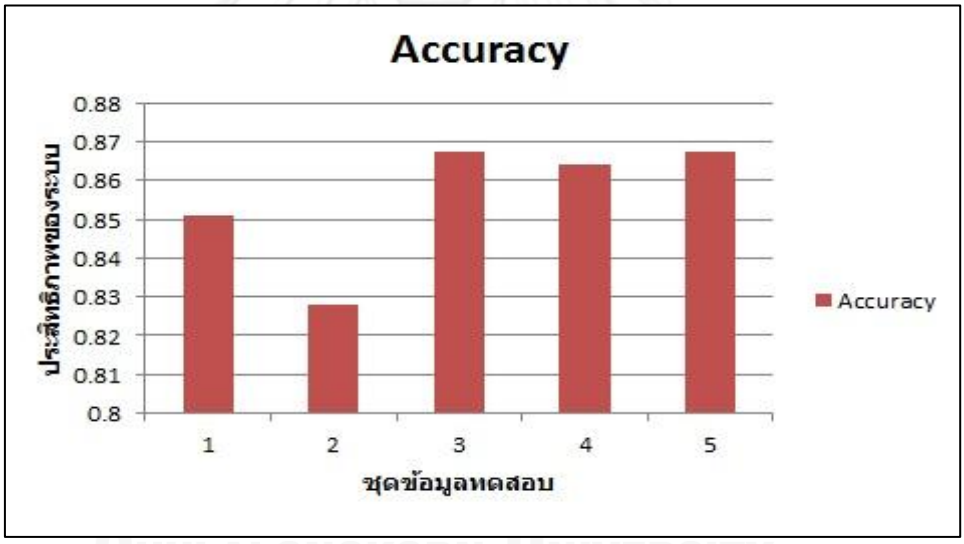
ภาพที่ 5.5 แสดงภาพผลการทดลองของระบบแนะนำการเชื่อมโยงกับเครื่องจักรเรียนรู้ Naïve Bayes

3. ผลการทดลองของระบบแนะนำการเชื่อมโยงกับเครื่องจักรเรียนรู้ SVM



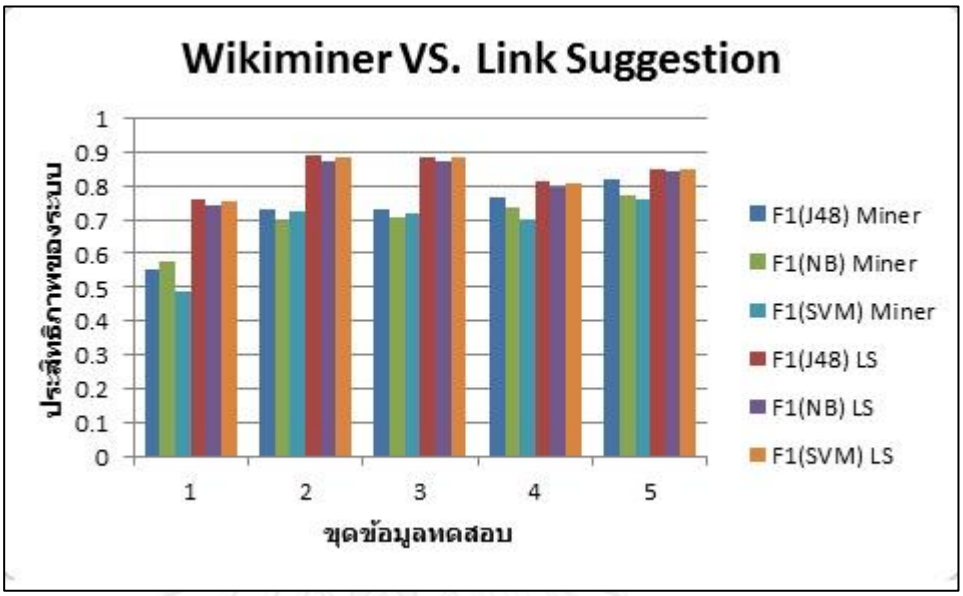
ภาพที่ 5.6 แสดงภาพผลการทดลองของระบบแนะนำการเชื่อมโยงกับเครื่องจักรเรียนรู้ SVM

4. ผลการทดลองของระบบแนะนำการเชื่อมโยงในส่วนแก้ไขปัญหาความกำกวม



ภาพที่ 5.7 แสดงภาพผลการทดลองของระบบแนะนำการเชื่อมโยงในส่วนแก้ไขปัญหาความกำกวม

5.2.3 เปรียบเทียบประสิทธิภาพระหว่างระบบวิกิไมเนอร์กับระบบแนะนำการเชื่อมโยง



ภาพที่ 5.8 แสดงภาพเปรียบเทียบประสิทธิภาพระหว่างระบบวิกิไมเนอร์กับระบบแนะนำการเชื่อมโยง

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 6
สรุปผลการวิจัยและข้อเสนอแนะ

6.1 สรุปผลการวิจัย

ในงานวิจัยนี้ เรานำเสนอกระบวนการแนะนำการเชื่อมโยงคำสำคัญในเอกสารข้อความไปยังเอกสารวิกิพีเดียปลายทาง โดยเราจะสนใจเฉพาะวิกิพีเดียภาษาไทยเป็นการเฉพาะ ขั้นตอนที่นำเสนอได้รวมเอาขั้นตอนการสกัดคำสำคัญแบบอัตโนมัติจากเอกสารนำเข้า และขั้นตอนการแก้ไขปัญหาคำกำกวม (คือจะต้องแนะนำการเชื่อมโยงไปยังบทความวิกิพีเดียที่เกี่ยวข้อง) เข้าไว้ในขั้นตอนเดียวกัน ระบบจะวิเคราะห์ และสกัดคุณลักษณะจากทุกๆ แคนดิเดตคำสำคัญที่ปรากฏในเอกสาร เพื่อนำไปใช้ฝึกสอนตัวจำแนกเครื่องจักรเรียนรู้เพื่อเรียนรู้วิธีที่จะแนะนำการเชื่อมโยงไปยังบทความวิกิพีเดียปลายทาง จากวิธีการดังกล่าว เราได้ทำการวัดประสิทธิภาพของระบบของเป็นสองขั้นตอนด้วยกัน โดยที่ ขั้นตอนแรก คือ การสกัดคำสำคัญแบบอัตโนมัติจากเอกสารนำเข้า เราได้ใช้ตัวจำแนกเครื่องจักรเรียนรู้เป็นตัวสกัดคำสำคัญและเลือกคำ เราได้ทำการทดสอบกับตัวจำแนกเครื่องจักรเรียนรู้ C4.5, Naivebayes และ Support Vector Machine จากข้อมูลฝึกสอนและทดสอบ ทั้ง 5 กลุ่ม ซึ่งเป็นบทความวิกิพีเดียภาษาไทย ระบบแนะนำการเชื่อมโยงสกัดคำและเลือกคำสำคัญได้ถูกต้องประมาณ 80% ในส่วนขั้นตอนที่ การแก้ไขปัญหาคำกำกวม ระบบแนะนำการเชื่อมโยง ได้ทำการเลือกบทความปลายทางให้กับคำสำคัญที่เลือกได้ถูกต้องประมาณ 85% นอกจากนี้เรายังได้นำข้อมูลบทความวิกิพีเดียภาษาไทย มาทำการทดสอบกับระบบวิกิไมเนอร์ เนื่องจากระบบวิกิไมเนอร์เมื่อทำการทดสอบกับบทความวิกิพีเดียภาษาอังกฤษ สกัดคำและเลือกคำสำคัญได้ถูกต้อง ประมาณ 70% ซึ่งเมื่อเราได้ทำการทดลองระบบวิกิไมเนอร์กับชุดข้อมูลฝึกสอนและทดสอบกับบทความวิกิพีเดียภาษาไทย ก็ยังคง สกัดคำและเลือกคำสำคัญได้ถูกต้องเพียง 70% จะเห็นได้ว่า ถ้าทำการทดลองกับข้อมูลบทความวิกิพีเดียภาษาไทยแล้ว ระบบแนะนำการเชื่อมโยงที่ได้นำเสนอมีประสิทธิภาพที่สูงกว่าระบบวิกิไมเนอร์

จากขั้นตอนกระบวนการที่นำเสนอนั้น สามารถสรุปกระบวนการทำงานได้เป็นสองส่วนหลักๆ ส่วนแรกคือการแก้ปัญหาของภาษาไทยที่มีลักษณะการเขียนที่แตกต่างจากภาษาอังกฤษ ตัวอย่างเช่น ไม่มีช่องว่างแยกระหว่างคำที่ชัดเจน ไม่มีการแปลงรูปกริยาเพื่อแสดงเวลา ไม่มีการใช้อักษรตัวใหญ่เพื่อระบุว่าเป็นคำนามเฉพาะนั้น เราได้ใช้ โปรแกรม Lexto มาช่วยในการตัดคำ และแก้ปัญหาในส่วนของภาษาทำให้เราได้คำสำคัญซึ่งมีความถูกต้องแม่นยำ และส่วนที่สองคือกระบวนการแนะนำการเชื่อมโยงคำสำคัญไปยังเอกสารวิกิพีเดีย เราได้นำเสนอชุดของคุณลักษณะที่มีความเหมาะสมในการเลือกคำสำคัญเพื่อที่จะสร้างการเชื่อมโยง จากกระบวนการซึ่งแยกได้เป็นสองส่วนที่ชัดเจนนี้ ทำให้เราได้เล็งเห็นการที่จะนำเอากระบวนการแนะนำการเชื่อมโยงคำสำคัญไปยังเอกสารวิกิพีเดียของงานวิจัยชิ้นนี้ไปใช้กับภาษาอื่นได้ เพียงแค่เปลี่ยนการทำงานในส่วนแรกคือเครื่องมือในการตัดคำหรือแบ่งคำ

จากขั้นตอนกระบวนการแนะนำการเชื่อมโยงคำสำคัญไปยังเอกสารวิกิพีเดีย จะเห็นได้เราว่าเราได้ใช้เครื่องจักรเรียนเป็นส่วนช่วยในการเลือกหรือระบุคำสำคัญที่จะใช้ในการแนะนำลิงก์ ซึ่งเครื่องจักรเรียนรู้นั้นจะทำงานได้อย่างมีประสิทธิภาพนั้นมีปัจจัยอยู่สองส่วนหลักๆด้วยกันคือ ชุด

คุณลักษณะที่นำมาสร้างเป็นโมเดลฝึกสอน และรูปแบบของโมเดลฝึกสอนที่เลือก โดยที่ทั้งสองส่วนที่กล่าวมานั้นจะต้องสามารถทำงานร่วนกันได้อย่างมีประสิทธิภาพ จากผลการทดสอบที่ได้นั้นเราจึงแนะนำว่า ชุดของคุณลักษณะที่เราเลือกใช้ในงานวิจัยชิ้นนี้คือ ค่าความน่าจะเป็นในการทำลิงก์ ค่าความคล้ายคลึงกันระหว่างเอกสาร ตำแหน่งที่ปรากฏของคำสำคัญ เป็นชื่อบทความหรือไม่ เป็นคำอักฤษหรือไม่ เป็นคำกำกวมหรือไม่และเป็นชื่อบทความเปลี่ยนทางหรือไม่ ส่วนรูปแบบของโมเดลฝึกสอนที่แนะนำคือ NaiveBayes เนื่องจากมีความแม่นยำใกล้เคียงกับโมเดลอื่นแล้วนั้นแต่สามารถทำงานได้รวดเร็วกว่า

ประสิทธิภาพการทำงานของเครื่องจักรเรียนรู้ในการเลือกคำสำคัญนั้นจากผลการทดลองที่ได้ในแต่ละกลุ่มของชุดข้อมูลที่นำมาฝึกสอนและทดสอบกระบวนการแนะนำการเชื่อมโยงของวิกิพีเดียภาษาไทยจะเห็นได้ว่าชุดข้อมูลในกลุ่มที่ มีสัดส่วนระหว่าง จำนวนลิงก์ของบทความกับคำสำคัญที่มากขึ้นจะทำให้ประสิทธิภาพของเครื่องจักรเรียนรู้นั้นสูงขึ้นตามไปด้วย จากข้อมูลดังกล่าว จึงขอสรุปลักษณะของบทความที่ควรที่จำใช้เป็นข้อมูลในฝึกสอนเครื่องจักรเรียนรู้นั้นควรที่จะมี สัดส่วนระหว่างลิงก์กับคำสำคัญมากกว่าหรือเท่ากับ 0.04 แต่ในส่วนของความถูกต้องในการเลือกบทความวิกิพีเดียปลายทางนั้น บทความในแต่ละกลุ่มนั้นไม่มีความแตกต่างอย่างเห็นได้ชัดจึงสรุปได้ว่า สัดส่วนระหว่างลิงก์กับคำสำคัญไม่มีผลต่อความถูกต้องในการเลือกบทความปลายทาง

6.2 ข้อจำกัด

ในการดำเนินงานวิจัยนี้มีข้อจำกัดในการใช้งานระบบดังต่อไปนี้

1. ในส่วนของการสกัดคำสำคัญ ในภาษาไทยนั้นต้องให้การตัดคำของโปรแกรมเล็กโตซึ่งถ้าต้องการนำระบบแนะนำการเชื่อโยงไปใช้กับภาษาอื่นๆ จะต้องเปลี่ยนเครื่องมือในการตัดคำ
2. ระบบแนะนำการเชื่อมโยง รองรับเฉพาะบทความที่อยู่ภายในฐานข้อมูลวิกิพีเดียภาษาไทยเท่านั้น จึงอาจทำให้คำสำคัญบางคำไม่ถูกแนะนำ
3. ในส่วนของระบบวิกิไมเนอร์ ข้อมูลวิกิพีเดียภาษาไทยไม่สามารถสกัดคุณลักษณะหมวดหมู่ จึงอาจทำให้ประสิทธิภาพที่ได้ไม่เต็มที่

6.3 แนวทางการวิจัยต่อไป

แต่อย่างไรก็ดีระบบที่นำเสนอ สามารถให้ค่าความแม่นยำ และค่าความระลึกในขั้นตอนการเลือกคำสำคัญประมาณ 80% ได้ผลลัพธ์ที่น่าพอใจเมื่อเทียบกับงานวิจัยต้นแบบของ [3], [4] เมื่อทำการทดลองกับข้อมูลบทความวิกิพีเดียภาษาไทย เรามีความเชื่อว่า เรายังคงพบช่องทางในการเพิ่มประสิทธิภาพของระบบได้อีก ตัวอย่างเช่น เราสามารถทดสอบกับตัวจำแนกเครื่องจักรเรียนรู้ชนิดอื่นๆ วิเคราะห์หาคุณลักษณะอื่นๆ เช่น คุณลักษณะทางภาษา ที่จะเอื้อประโยชน์ให้กับตัวจำแนก ทดสอบใช้ตัวจำแนกหลายๆ ตัวร่วมกัน เพิ่มขั้นตอนการวิเคราะห์ความคล้ายคลึงเชิงความหมายในแต่ละย่อหน้า หรือแม้กระทั้ง ดูช่วงเวลาที่บทความนั้นกำลังเป็นที่กล่าวถึงหรือไม่ ซึ่งจากการวิเคราะห์ดังกล่าว เราเล็งเห็นว่า น่าจะช่วยเพิ่มประสิทธิภาพของระบบแล้วยังทำให้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ระบบทำงานได้ละเอียดและแม่นยำยิ่งขึ้น นอกจากนี้เรายังวางแผนที่จะทำการศึกษาผลลัพธ์ที่ได้ผ่านจากผู้ใช้งานระบบ (user study) ที่เรานำเสนออีกด้วย

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

รายการอ้างอิง

1. วิกิพีเดีย สารานุกรม,. Available from: http://th.wikipedia.org/wiki/หน้าหลัก.
2. Kulkarni, S., et al., *Collective annotation of Wikipedia entities in web text*, in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009, ACM: Paris, France. p. 457-466.
3. Mihalcea, R. and A. Csomai, *Wikify!: linking documents to encyclopedic knowledge*, in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 2007, ACM: Lisbon, Portugal. p. 233-242.
4. Milne, D. and I.H. Witten, *Learning to link with wikipedia*, in *Proceedings of the 17th ACM conference on Information and knowledge management*. 2008, ACM: Napa Valley, California, USA. p. 509-518.
5. Ratinov, L., et al., *Local and global algorithms for disambiguation to Wikipedia*, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. 2011, Association for Computational Linguistics: Portland, Oregon. p. 1375-1384.
6. Ngo, T., *Data mining: practical machine learning tools and technique, third edition by Ian H. Witten, Eibe Frank, Mark A. Hell.* SIGSOFT Softw. Eng. Notes, 2011. **36**(5): p. 51-52.
7. *LexTo*. Available from: http://www.sansarn.com/lexto/.
8. บทความสอนการใช้งานวิกิพีเดีย : วิกิพีเดียลิงก์,. Available from: http://th.wikipedia.org/wiki/วิกิพีเดีย:สอนการใช้งาน_(วิกิพีเดียลิงก์).
9. Kleinberg, J.M., *Authoritative sources in a hyperlinked environment.* J. ACM, 1999. **46**(5): p. 604-632.
10. Horowitz, D. and S.D. Kamvar, *The anatomy of a large-scale social search engine*, in *Proceedings of the 19th international conference on World wide web*. 2010, ACM: Raleigh, North Carolina, USA. p. 431-440.
11. Abellán, J. and A.R. Masegosa. *Split Criterions for Variable Selection Using Decision Trees*. in *Proceedings of the 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. 2007.
12. Wu, H.C., et al., *Interpreting TF-IDF term weights as making relevance decisions.* ACM Trans. Inf. Syst., 2008. **26**(3): p. 1-37.
13. Salton, G. and M.J. McGill, *Introduction to Modern Information Retrieval*. 1986: McGraw-Hill, Inc. 400.
14. Quinlan, J.R., *C4.5: programs for machine learning*. 1993: Morgan Kaufmann Publishers Inc. 302.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ภาคผนวก

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ประวัติผู้เขียนวิทยานิพนธ์

นายสมภพ เซี่ยงคิ้ว เกิดเมื่อวันที่ 28 มกราคม พ.ศ. 2528 ที่จังหวัดสิงห์บุรี สำเร็จการศึกษาหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยกรุงเทพ และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ที่ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2552

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University