เทคนิคการค้นคืนสารสนเทศข้ามภาษา (ไทย–อังกฤษ) โดยใช้ออนโทโลยี

Ontology-based Technique for Cross-Language (Thai-English) Information Retrieval

ฉัตรชัย อินทรประพันธ์ (Chatchai Inparaprapana)* และ ไกรศักดิ์ เกษร (Kraisak Kesorn)*

บทคัดย่อ

ในยุคแห่งระบบข้อมูลสารสนเทศ เอกสารต่างๆ ได้กระจายอยู่ทั่วไปบนระบบเครือข่ายอินเทอร์เน็ตและมีแนวโน้มที่จะเพิ่มมากขึ้นอย่างรวดเร็วต่อไปในอนาคต เอกสารเหล่านี้ถูกเขียนขึ้นโดยใช้ภาษาต่างๆ กัน ถึงแม้ว่าระบบค้นคืนสารสนเทศ (Information Retrieval) ในปัจจุบันจะมีประสิทธิภาพสูงมากในการค้นคืนข้อมูล อย่างไรก็ตาม ระบบค้นคืนสารสนเทศเหล่านี้จะค้นหาเฉพาะเอกสารที่ถูกเขียนโดยใช้ภาษาเดียวกับคำสำคัญ (Keywords) เท่านั้น ดังนั้นงานวิจัยนี้ได้นำเสนอวิธีการสืบค้นสารสนเทศข้ามภาษาสำหรับเอกสารภาษาไทยและภาษาอังกฤษ แนวคิดสำคัญของงานวิจัยนี้ประกอบด้วย 2 ส่วนหลัก คือ 1) การใช้ออนโทโลยีในการทำดัชนีเอกสารเพื่อรองรับการค้นคืนข้อมูลข้ามภาษาและใช้โครงสร้างของออนโทโลยีช่วยในการค้นคืนข้อมูลเชิงความหมาย (Semantic Search) และ 2) พัฒนาวิธีการในการจัดเรียงข้อมูล (Ranking) ผลลัพธ์ที่ได้จากการค้นคืนข้อมูลสำหรับระบบค้นหาข้ามภาษา ผลการทดลองแสดงให้เห็นว่าระบบสามารถค้นหาเอกสารได้ทั้งเอกสารภาษาไทยและภาษาอังกฤษโดยใช้คำสำคัญเพียงภาษาเดียวเท่านั้นและสามารถปรับปรุงค่ารีคอล (Recall) และค่าพรีซีชั่น (Precision) ให้สูงขึ้น

คำสำคัญ: ออนโทโลยี การค้นคืนสารสนเทศข้ามภาษา การจัดลำดับข้อมูล การค้นคืนสารสนเทศเชิงความหมาย

Abstract

There are billions of documents distributed in the Internet and these numbers trend to increase dramatically in the future. These documents are written using several languages. Although the existing search engines obtain high retrieval performance, the problem is the Information Retrieval System will retrieve only desire documents written by the similar language in the query. In other words, the search engine is unable to find relevant documents written by different languages from the query. This paper presents an ontology-base technique for cross-language information retrieval (CLIR). The novelty of this paper are 1) using an ontology to store information in the hierarchical structure in order to provide semantic search; 2) ranking technique of results of cross-language information retrieval by modifying cosine similarity formula by taking a language weight into account. The experiment results show that the proposed technique allows the system search relevant documents that are written a different language from the query and, thus, it can improve the precision and recall significantly.

Keyword: Ontology model, Cross-Language Information Retrieval, Ranking, Semantic search.

* *ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร*

1. บทนำ

ในปัจจุบัน การเก็บข้อมูลหรือเอกสารในรูปแบบของสื่ออิเล็กทรอนิกส์มีมากขึ้น ซึ่งข้อมูลและเอกสารเหล่านี้ได้กระจายอยู่ทั่วไปในระบบเครือข่ายอินเทอร์เน็ต ระบบเครือข่ายอินเทอร์เน็ตในปัจจุบันได้พัฒนาขึ้นอย่างรวดเร็ว ทำให้สามารถเข้าถึงข้อมูลหรือเอกสารได้ง่ายและสะดวกมากขึ้น ข้อมูลหรือเอกสารต่างๆ ที่กระจายอยู่ทั่วไปนั้น ถูกเขียนขึ้นโดยใช้ภาษาที่หลากหลายแตกต่างกัน เช่น ภาษาอังกฤษ ภาษาไทย ภาษาญี่ปุ่น เป็นต้น ข้อดีของข้อมูลที่มีหลากหลายภาษานั้นคือผู้ใช้มีข้อมูลที่หลากหลาย ซึ่งสามารถนำข้อมูลเหล่านี้มาตรวจสอบยืนยันความถูกต้องได้ และสามารถนำไปใช้ประโยชน์ด้านอื่นๆ ได้อีกไม่ว่าจะเป็นด้านการศึกษาหรือด้านการทำงานทั่วไป อย่างไรก็ตามข้อจำกัดที่สำคัญคือปัจจุบันผู้ใช้ต้องทำการค้นหาข้อมูลโดยใช้คำสำคัญ (Keyword) ที่เป็นภาษาเดียวกับเอกสารเท่านั้น ตัวอย่างเช่น ผู้ใช้ไม่สามารถที่จะค้นหาเอกสารภาษาอังกฤษโดยใช้คำสำคัญเป็นภาษาไทยได้ นอกจากนี้ผู้ใช้อาจจะไม่รู้จะใช้คำศัพท์ใดที่จะอธิบายถึงข้อมูลที่ตนเองกำลังค้นหาได้ดีที่สุด ซึ่งมีผลทำให้ระบบการค้นคืนสารสนเทศไม่สามารถค้นหาเอกสารที่ผู้ใช้ต้องการได้อย่างถูกต้องแม่นยำเท่าที่ควร สาเหตุสำคัญของปัญหานี้ คือ ผู้ใช้มีความรู้จำกัดเกี่ยวกับคำศัพท์ต่างๆ ที่ไม่ใช่ภาษาประจำชาติของตนและความหมายของคำศัพท์ในภาษาต่างๆ มีความไม่แน่นอนตายตัว ซึ่งจากข้อจำกัดนี้จะพบว่าสาเหตุมาจาก คำหนึ่งคำสามารถมีหลายความหมาย (Polysemy) เช่น apple อาจจะหมายถึงผลไม้ชนิดหนึ่งหรือเครื่องคอมพิวเตอร์ยี่ห้อหนึ่ง และคำหลายคำสามารถมีความหมายเดียวกัน (Synonym) เช่น image และ photo ทั้งสองคำต่างก็มีความหมายว่ารูปภาพ ระบบค้นคืนสารสนเทศในปัจจุบันยังมีประสิทธิภาพที่ไม่สูงมากนักในการแก้ไขปัญหาเหล่านี้ และยังไม่สามารถเข้าใจถึงความหมายที่แท้จริงของคิวรี่ (Query) นอกจากนี้ระบบค้นคืนสารสนเทศบนอินเทอร์เน็ตในปัจจุบัน เช่น Google Yahoo ThaiQuest ยังไม่สามารถค้นคืนสารสนเทศข้ามภาษา (Cross-Language) ได้ ผู้ใช้ใส่คำสำคัญที่ต้องการค้นหาเป็นภาษาไทย ระบบค้นหา เช่น Google จะทำการค้นหาเฉพาะเอกสารที่เป็นภาษาไทยให้ผู้ใช้เท่านั้น แต่ในความเป็นจริงอาจจะมีเอกสารอื่นๆ อีกมากมายที่เกี่ยวข้องแต่ถูกเขียนเป็นภาษาอังกฤษอยู่บนระบบอินเทอร์เน็ต แต่ระบบค้นคืนสารสนเทศที่มีอยู่ไม่สามารถหาเอกสารเหล่านั้นได้ ทั้งนี้เนื่องจากไม่มีกลไกที่สนับสนุนการค้นหาเอกสารข้ามภาษานั่นเอง นอกจากว่าในแท็ก (Tag) ต่างๆ ของเอกสารจะเขียนขึ้นโดยระบุทั้งสองภาษา เช่น “CarOnline.net | นิตยสารรถยนต์ออนไลน์” แต่วิธีนี้ยังไม่ถือเป็นระบบค้นคืนสารสนเทศข้ามภาษาอย่างแท้จริง เนื่องจากมีการระบุข้อมูลไว้ในแท็กต่างๆ โดยคน (Manual) ซึ่งต้องใช้แรงงานสูงและเสียเวลา ดังนั้นงานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาระบบค้นคืนสารสนเทศข้ามภาษาขึ้น เพื่อช่วยให้ผู้ใช้สามารถค้นคืนข้อมูลได้ทั้งภาษาไทยและภาษาอังกฤษโดยใช้คำสำคัญ (Keyword) ในการค้นคืนเพียงแค่ครั้งเดียว

ในบทความนี้ประกอบด้วยส่วนต่างๆ ดังนี้ ส่วนที่ 2 เป็นงานวิจัยที่เกี่ยวข้อง ส่วนที่ 3 อธิบายถึงการทำงานของระบบค้นหาข้ามภาษาที่นำเสนอ ส่วนที่ 4 การทดลองและการวิเคราะห์ผลการทดลอง และส่วนที่ 5 เป็นการสรุปงานวิจัย

2. งานวิจัยที่เกี่ยวข้อง

จากการศึกษาในงานวิจัยที่เกี่ยวข้องกับพัฒนาระบบค้นคืนสารสนเทศนั้น ได้มีนักวิจัยหลายท่านได้พัฒนาระบบขึ้นมาโดยใช้เทคนิคที่แตกต่างกัน เช่น Neural Networks, Longest-matching, Query Expansion เป็นต้น ในประเทศไทยนั้น งานวิจัยที่เกี่ยวกับการพัฒนาระบบค้นหาข้ามภาษาระหว่างภาษาไทยและภาษาอังกฤษมีจำนวนน้อย ในหัวข้อนี้จะให้ความสำคัญกับการศึกษางานวิจัยที่เกี่ยวข้องกับการค้นคืนสารสนเทศข้ามภาษา (CLIR: Cross-Language Information Retrieval) และจัดลำดับผลลัพธ์การค้นหาข้อมูล (Ranking) ของระบบค้นคืนสารสนเทศข้ามภาษาเท่านั้น

2.1 การค้นคืนสารสนเทศข้ามภาษา

แนวคิดเกี่ยวกับการค้นคืนสารสนเทศข้ามภาษานี้ถูกพัฒนาขึ้นเมื่อปี ค.ศ. 1964 ซึ่งในระยะแรกของการพัฒนานั้น จะพัฒนาโดยใช้เทคนิคของการใช้พจนานุกรม (Thesaurus) มาช่วย แต่การใช้เทคนิคมีข้อจำกัดอยู่คือ ผู้ใช้ต้องมีความชำนาญสูงเพราะจะต้องใช้คำศัพท์ที่เฉพาะเท่านั้น เพราะฉะนั้นจึงยากต่อการใช้งานสำหรับคนทั่วไปที่ไม่รู้คำศัพท์เฉพาะ นอกจากนั้นยังมีจำนวนคำศัพท์ตายตัวและการเพิ่มคำศัพท์ลงไปภายหลังทำได้ยากเพราะว่าแต่ละภาษาจะมีคำศัพท์ใหม่ๆ เพิ่มขึ้นตลอดเวลา หลังจากนั้นได้มีการพัฒนาเทคนิคต่างๆ ที่เกี่ยวข้องกับการค้นคืน

สารสนเทศข้ามภาษาขึ้น เช่น ทัศนีย์ ศูนย์กลาง และคณะ [1] ได้พัฒนาระบบค้นคืนสารสนเทศระหว่างภาษาไทยและภาษาอังกฤษ โดยมีขอบเขตอยู่ที่คำทับศัพท์ทั้งภาษาไทยและอังกฤษ งานวิจัยนี้ใช้เทคนิคโครงข่ายประสาทเทียม (Neural Networks) ซึ่งใช้สำหรับเรียนรู้แปลงคำภาษาไทยและภาษาอังกฤษเป็นคำอ่าน ระบบทำการค้นคืนข้อมูลโดยใช้เทคนิคการเปรียบเทียบสายอักขระ (String Matching) โดยการค้นคืนแบบเปรียบเทียบอักขระจะเป็นค้นคืนเอกสารที่มีสายอักขระเหมือนกันเท่านั้น เอกสารที่ไม่มีสายอักขระที่ไม่เหมือนกันจะถูกคัดออก ทำให้ประสิทธิภาพการค้นคืนข้อมูลต่ำ สำหรับภาษาจีน Chen [2] ได้พัฒนาระบบค้นคืนข้อมูลข้ามภาษาระหว่างภาษาจีนและภาษาอังกฤษโดยใช้เทคนิค Longest-matching ซึ่งจะเป็นการเทียบสายอักขระที่ยาวที่สุดกับพจนานุกรม ข้อจำกัดคือวิธีนี้ยังไม่สามารถค้นคืนข้อมูลเชิงความหมายได้ งานวิจัยใน [3] [4] [5] ได้พัฒนาระบบค้นคืนข้อมูลข้ามภาษา โดยใช้เทคนิค Query Expansion ซึ่งเป็นวิธีในการขยายคิวรี่หรือเพิ่มศัพท์ในการค้นคืน ซึ่งจะใช้พจนานุกรมและบริการแปลภาษาออนไลน์มาช่วยให้การแปลและเพิ่มศัพท์ ซึ่งทำให้ได้คำศัพท์ที่มีความหมายเดียวกันเพิ่มขึ้นและใช้ในการค้นคืนได้ ข้อกำจัดของเทคนิคนี้คือยังไม่สามารถค้นคืนเอกสารเชิงความหมายได้ และเอกสารที่ได้ออกมานั้นอาจจะไม่ตรงกับความต้องการ เนื่องจากในการแปลงตัวอักษรนั้นจะต้องแปลงตัวอักษรของภาษาฝรั่งเศสบางตัวในอยู่ในรูปแบบของตัวอักษรภาษาอังกฤษ เช่น ‘a’ แทนด้วย ‘a’ ซึ่งทำให้การขยายคิวรี่และค้นหามีความผิดพลาดไปบ้าง Çelebi และคณะ [6] และ Zhou และคณะ [7] ได้พัฒนาระบบค้นคืนสารสนเทศข้ามภาษาสำหรับภาษาตุรกี-ภาษาอังกฤษ และภาษาฝรั่งเศส-ภาษาอังกฤษตามลำดับ โดยงานวิจัยทั้ง 2 งานได้ใช้เทคนิค LSI (Latent Semantic Indexing) ซึ่งเทคนิคนี้จะพิจารณาเอกสารใดๆ ที่มีจำนวนคำสำคัญปรากฏร่วมกันสูงจะถือว่าเอกสารเหล่านั้นมีความสัมพันธ์กันเชิงความหมาย (Semantically Related) และเอกสารใดๆ ที่มีคำสำคัญปรากฏร่วมกันอยู่จำนวนน้อยจะถือว่ามีความสัมพันธ์กันเชิงความหมายต่ำ อย่างไรก็ตามข้อเสียของเทคนิคนี้คือ บางเอกสารที่สำคัญแต่ปรากฏว่าพบคำสำคัญน้อยจะถูกตัดสินว่ามีความสัมพันธ์กันต่ำ ทำให้ผลลัพธ์ที่ได้จากการค้นคืนสารสนเทศมีประสิทธิภาพไม่สูงมากนัก Embley และคณะ [8] ได้พัฒนาระบบค้นคืนสารสนเทศข้ามภาษาระหว่างภาษาญี่ปุ่นและภาษาอังกฤษ โดยการใช้ออนโทโลยีเข้ามาช่วยในการเก็บข้อมูลและแทนข้อมูล (Represent data) ความสัมพันธ์ของคอนเซฟท์ในออนโทโลยีสามารถช่วยให้ระบบดังกล่าวสามารถค้นหาเชิงความหมายได้ (Semantic search) หมายความว่าระบบสามารถค้นหาเอกสารที่เกี่ยวข้องได้โดยที่เอกสารนั้นไม่มีคำสำคัญที่ใช้ค้นหาปรากฏอยู่เลยก็ตาม

2.2 การจัดลำดับผลลัพธ์จากการค้นคืนสารสนเทศ

นักวิจัยหลายท่านได้มุ่งพัฒนาเทคนิคต่างๆ เพื่อให้ระบบค้นคืนสารสนเทศได้ผลลัพธ์ที่ได้ออกมามีความถูกต้องและตรงกับความต้องการของผู้ใช้มากที่สุด อย่างไรก็ตามส่วนที่สำคัญอีกส่วนหนึ่งคือการจัดลำดับผลลัพธ์การค้นคืนข้อมูล (Ranking) เนื่องจากจะช่วยทำให้ผู้ใช้สามารถพบข้อมูลที่กำลังค้นหาได้เร็วยิ่งขึ้น Wei และคณะ [9] วิจัยในเรื่องการจัดลำดับเอกสารที่ถูกเขียนโดยภาษาที่ต่างกันโดยใช้อัลกอริทึม L2R [10] ซึ่งเป็นอัลกอริทึมที่นิยมใช้ในการจัดลำดับของระบบค้นคืนข้ามภาษา ผลที่ได้จากการทดลองนั้นได้นำมาเปรียบเทียบกับวิธี SVM ซึ่งวิธี L2R จะสามารถจัดลำดับได้มีถูกต้องกว่าวิธี SVM Zhou และ Wade [5] ได้พัฒนาระบบค้นคืนสารสนเทศข้ามภาษาระหว่างภาษาจีนและภาษาอังกฤษและพัฒนาวิธีการจัดลำดับของผลลัพธ์ใหม่โดยใช้เทคนิค Latent Dirichlet Allocation (LDA) [11] ผลการทดลองพบว่าผลลัพธ์ที่ได้จากการจัดลำดับใหม่นั้นมีการเปลี่ยนแปลงเพียงเล็กน้อย และการจัดลำดับนี้จะไม่จัดลำดับใหม่ทั้งหมดเพียงแต่จัดลำดับใหม่ใน 50 อันดับแรกเท่านั้น เมื่อผลลัพธ์ที่ได้มามีจำนวนมากขึ้นพบว่าผลลัพธ์ที่ตรงกับความต้องการไม่ได้ถูกจัดลำดับมาอยู่ในลำดับต้นๆ Cao และคณะ [12] ได้ใช้เทคนิค SVM [13] มาช่วยในการจัดลำดับผลลัพธ์ ซึ่งนักวิจัยได้นำเทคนิคนี้รวมเข้าระบบค้นคืนสารสนเทศที่พัฒนาขึ้นซึ่งทำให้ประสิทธิภาพในการค้นหาดีขึ้นมากกว่าระบบเดิม เพราะว่าเทคนิคนี้ได้มีรวมค่าน้ำหนักและสร้างแบบจำลองเชิงเส้นขึ้นเพื่อคำนวณครั้งคะแนนสุดท้ายก่อนส่งคืนเอกสารสู่ผู้ใช้ Zhou และคณะ [7] ได้ใช้เทคนิค LDA ได้ทำการทดลองในระบบค้นคืนข้ามภาษาระหว่างภาษาฝรั่งเศส-อังกฤษ ในการทดลองได้ทำการเปรียบเทียบระหว่างการจัดลำดับแบบ LDA และ LSI ซึ่งจากผลการทดลองนั้นเทคนิค LDA สามารถจัดลำดับได้ดีกว่า เพราะว่าเทคนิค LDA นี้จะจัดลำดับให้เอกสารที่มีคำที่

เกี่ยวข้องกับคำสำคัญที่มีจำนวนมากกว่ามาอยู่ในลำดับแรก แต่เทคนิค LSI จะจัดลำดับให้เอกสารเฉพาะที่มีคำสำคัญจำนวนมากมาอยู่ในลำดับแรก

จากการศึกษาและเปรียบเทียบงานวิจัยที่เกี่ยวข้องสามารถสรุปได้ดังตารางที่ 1

ตารางที่ 1 *สรุปงานวิจัยที่เกี่ยวข้องกับการค้นคืนสารสนเทศข้ามภาษา*

Author	Technique	String Machining Search	Semantic Search	Cross-language Ranking
[1]	Neural Network	✓	✘	✘
[2]	Longest Matching	✓	✘	✘
[3], [4]	Query Expansion	✓	✘	✘
[6]	LSI	✓	P	✘
[7]	LSI	✓	P	N/A
[5]	Query Expansion	✓	✘	N/A
[8]	Ontology	✓	✘	✘

✓ = มี ✘ = ไม่มี P = รองรับการทำงานได้บางส่วน (Partially) N/A = ไม่มีข้อมูลระบุชัดเจน

จากการศึกษาในงานวิจัยที่เกี่ยวข้องกับพัฒนาระบบค้นคืนสารสนเทศข้ามภาษาพบว่าระบบที่มีอยู่ในปัจจุบันยังมีข้อจำกัดสรุปได้ดังต่อไปนี้

1) ระบบที่มีอยู่มีการทำงานโดยใช้การเปรียบเทียบอักขระซึ่งทำให้การค้นหาสารสนเทศมีประสิทธิภาพต่ำ

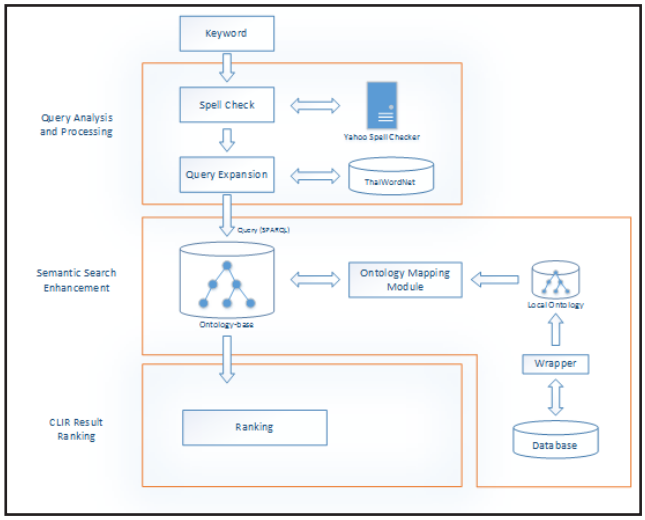
2) ระบบส่วนใหญ่ไม่รองรับการค้นคืนสารสนเทศเชิงความหมายหรือรองรับการค้นคืนข้อมูลเชิงความหมายได้เพียงบางส่วน (P) เท่านั้น (ยกเว้น [8]) เนื่องจากระบบดังกล่าวทำงานบนพื้นฐานของ LSI ซึ่งเป็นคำนวณทางสถิติ ดังนั้นประสิทธิภาพการทำงานจะไม่สูงเท่าออนโทโลยีซึ่งมีการระบุความสัมพันธ์ของคอนเซฟท์ต่างๆ ไว้ชัดเจนและถูกต้องมากกว่า

3) จากการศึกษางานวิจัยเดิมพบว่างานวิจัยส่วนใหญ่ยังไม่ให้ความสำคัญการจัดลำดับผลลัพธ์การค้นหาข้อมูลของระบบค้นคืนสารสนเทศข้ามภาษาโดยมุ่งเน้นแต่ประสิทธิภาพการค้นคืนสารสนเทศเท่านั้น

ด้วยความสำคัญของประเด็นปัญหาดังกล่าว ผู้วิจัยจึงเกิดแนวคิดที่จะพัฒนาระบบค้นหาข้ามภาษาสำหรับภาษาไทยและภาษาอังกฤษโดยใช้ออนโทโลยีเป็นฐานความรู้เพื่ออนุญาตให้ระบบสามารถค้นคืนสารสนเทศเชิงความหมายได้ และพัฒนาวิธีการเรียงลำดับผลลัพธ์การค้นหาข้อมูลให้มีประสิทธิภาพมากยิ่งขึ้น โดยกรอบแนวคิดของระบบ (Framework) ของงานวิจัยนี้ได้นำเสนอในหัวข้อถัดไป

3. กรอบการทำงานของระบบค้นคืนข้ามภาษา

หัวข้อนี้จะอธิบายถึงกรอบการทำงานของระบบค้นคืนข้ามภาษาสำหรับภาษาไทยและภาษาอังกฤษ แสดงดังภาพที่ 1 ประกอบด้วย 3 ส่วนหลักๆ ได้แก่ การวิเคราะห์และประมวลผลคิวรี่ (Query Analysis and Processing) การเพิ่มประสิทธิภาพการค้นคืนสารสนเทศเชิงความหมาย (Semantic Search Enhancement) และการเรียงลำดับผลลัพธ์สำหรับระบบค้นคืนสารสนเทศข้ามภาษา (CLIR Result Ranking)



ภาพที่ 1 *กรอบการทำงานของระบบค้นคืนสารสนเทศข้ามภาษาสำหรับภาษาไทย-อังกฤษ*

3.1 การวิเคราะห์และประมวลผลคิวรี่ (Query analysis and processing)

ส่วนแรกนี้เป็นกระบวนการประมวลผลคิวรี่จากผู้ใช้ เมื่อผู้ใช้ได้ป้อนคำสำคัญ (Keyword) ผ่านทางระบบค้นคืนสารสนเทศข้ามภาษา ระบบจะนำคำสำคัญที่ป้อนไว้มาวิเคราะห์และประมวลผล ในขั้นตอนแรกของส่วนนี้นั้นจะทำการตรวจสอบความถูกต้องของคำสำคัญที่ผู้ใช้ป้อนเข้ามา ในการตรวจสอบความถูกต้องนี้ผู้วิจัยไม่ได้พัฒนา

ระบบตรวจสอบคำผิดขึ้นมาเอง เนื่องจากไม่ใช่วัตถุประสงค์หลักของงานวิจัยและยังมีเครื่องมือที่มีประสิทธิภาพสามารถนำมาใช้งานได้ ในงานวิจัยนี้ผู้วิจัยใช้ "Yahoo Spell Checker API" เนื่องจากบริษัท Yahoo ได้จัด API นี้ให้นักพัฒนาระบบสำหรับ API ของ Google นั้นต้องเสียค่าใช้จ่าย หน้าที่หลักของ Yahoo Spell Checker API คือการตรวจสอบการสะกดคำว่าถูกหรือผิด หากผู้ใช้สะกดคำสำคัญผิดระบบจะแก้ให้อัตโนมัติและแสดงให้ผู้ใช้ทราบดังภาพที่ 2 ข้อดีของขั้นตอนนี้คือส่งเสริม (Enhance) ให้ผู้ใช้ใส่คำได้ถูกต้องและลดข้อผิดพลาดจากผู้ใช้ (Minimize errors) และส่งผลให้ระบบใช้คำสำคัญในการค้นหาได้ถูกต้องมากยิ่งขึ้น

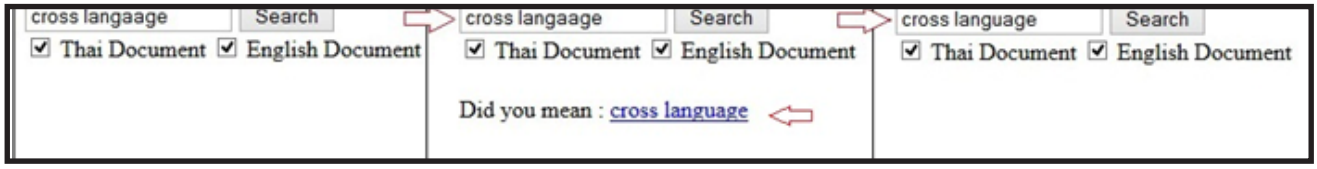
ในขั้นตอนถัดไประบบจะทำการขยายคำสำคัญในคิวรี่ (Query Expansion) หมายความว่าระบบจะนำคำสำคัญมาหาคำอื่นๆ ที่เกี่ยวข้อง (Relevant) ทั้งภาษาไทยและภาษาอังกฤษเพื่อรองรับปัญหา Synonym การขยายคำสำคัญในงานวิจัยนี้ทำงานบนพื้นฐานของ ThaiWordNet [14] โดย ThaiWordNet คือฐานข้อมูลที่รวบรวมคำศัพท์ภาษาไทยและภาษาอังกฤษไว้ ข้อมูลที่ได้จาก ThaiWordNet สามารถกำหนดได้ว่าจะให้เป็นข้อมูลที่อยู่ในรูปของ JSON หรือ XML ก็ได้ ซึ่งทั้งสองรูปแบบนี้เป็นรูปแบบของข้อมูลสำหรับแลกเปลี่ยนข้อมูลทางอิเล็กทรอนิกส์ ตัวอย่างของข้อมูลใน ThaiWordNet แสดงดังภาพที่ 3 ซึ่งอยู่ในรูปแบบของ JSON การขยายคำสำคัญในคิวรี่จะทำให้ได้คำสำคัญอื่นๆ ที่เกี่ยวข้องเพิ่มขึ้นและถูกใช้สำหรับค้นหาเอกสารที่เกี่ยวข้อง

เมื่อได้ผลลัพธ์กลับมาระบบจะต้องมีการจัดเรียงลำดับผลลัพธ์เหล่านี้โดยเรียงลำดับเอกสารที่ผู้ใช้น่าจะสนใจมากที่สุดเป็นอันดับแรก และลดลงไปเรื่อยๆ ในการคำนวณเพื่อเรียงลำดับข้อมูลดังกล่าวระบบจะต้องมีการกำหนดน้ำหนัก (Weight) ของคำสำคัญแต่ละคำเพื่อให้การจัดลำดับผลลัพธ์จากการค้นหาข้อมูลให้มีประสิทธิภาพมากยิ่งขึ้น โดยให้คำสำคัญเดิมที่ผู้ใช้ได้ป้อนมีน้ำหนักมากที่สุดคือ 1 และคำสำคัญอื่นๆ ที่ถูกขยายเพิ่มขึ้นจะมีน้ำหนักลดลงลำดับ

```
<search_word>image</search_word>
<rows>11</rows>
<data>
<itemid="1">
        <sense_id>21699</sense_id>
        <synset_offset>03931044</synset_offset>
        <ss_type>n</ss_type>
        <lex_filenum>06</lex_filenum>
        <lex_filename>noun.artifact</lex_filename>
        <words>picture, image, icon, ikon</words>
        <translate>ภาพ</translate>
</item>
<item id="2">
        <sense_id>17763</sense_id>
        <synset_offset>03265874</synset_offset>
        <ss_type>n</ss_type>
        <lex_filenum>06</lex_filenum>
        <lex_filename>noun.artifact</lex_filename>
        <words>effigy, image, simulacrum</words>
        <translate>หุ่น, หุ่นจำลอง</translate>
</item>
```

ภาพที่ 3 *ตัวอย่างข้อมูลที่ได้จาก ThaiWordnet*

ละ 0.2 ตามข้อมูลที่ได้จาก ThaiWordNet เพราะว่าจากข้อมูลที่ได้จาก ThaiWordNet นั้นคำที่ถูกจัดอยู่ในลำดับแรกๆ นั้นจะมีความสัมพันธ์กับคำสำคัญมาก และในลำดับถัดไปความสัมพันธ์ก็จะลดลงเรื่อยๆ ดังแสดงในภาพที่ 3 จะเห็นว่าคำที่อยู่ในลำดับ (id) ที่ 2 คำที่ได้มีเริ่มจะแตกต่างไปจากลำดับที่ 1 และหาก ThaiWordNet ได้ให้ข้อมูลมากกว่า 5 ลำดับ ระบบเองจะไม่นำคำนั้นมาใช้ เนื่องจากยิ่งลำดับมากขึ้นความหมายที่ได้ก็จะหากไกลกันมาก โดยจะมีวิธีในการขยายคีย์เวิร์ด จะแสดงเป็นขั้นตอนการทำงานอัลกอริทึม 1 โดยขั้นตอนการขยายคำสำคัญนั้นจะใช้ ThaiWordNet เข้ามาช่วยในการขยาย ซึ่ง ThaiWordNet นี้ เป็นเว็บเซอร์วิสที่จะส่งค่ากลับมาให้เป็นแบบ XML หรือ JSON ซึ่งสามารถใช้งานได้ผ่าน "http://th.asianwordnet.org /services/dictionary/json/en2th/keyword" และทำการแปลงข้อมูลที่ได้กลับมาอยู่ในรูปแบบของอาเรย์ เพื่อจะได้นำไปใช้ในการค้นหาได้ต่อไป



ภาพที่ 2 *ตัวอย่างการตรวจสอบคำสำคัญโดยใช้ Yahoo Spell Checker*

```
1 :  Input : keyword
2 :  url = “http://th.asianwordnet.org/services/dictionary/json/
en2th/” + keyword
3 :  data_json = get json from url
4 :  data_array = encode data_json to Array
5 :  foreach(data_array as data)
6 :      words = data[words]
7 :      translate = data[translate]
8 :      other_keyword = array
9 :      other_keyword = split word to Array
10 :     other_keyword = split translate to Array
11 : end foreach
12:  query_keyword = function array_unique
```

อัลกอริทึม 1 ขั้นตอนการทำงานการขยายคำสำคัญโดยใช้ ThaiWordnet

3.2 การเพิ่มประสิทธิภาพการค้นคืนสารสนเทศเชิงความหมาย (Semantic Search Enhancement)

ส่วนนี้เป็นส่วนของการค้นคืนเอกสารที่สัมพันธ์กันเชิงความหมาย (Semantic Rated) ระบบค้นคืนข้ามภาษาจะนำคำสำคัญที่ได้จากการขยายคิวรี่ (Query Expansion) ในส่วนที่ 1 ไปทำการค้นคืนข้อมูลในออนโทยีที่ได้ออกแบบไว้ ซึ่งได้อธิบายความสัมพันธ์ของข้อมูลในรูปแบบของ OWL (Web Ontology Language) ซึ่งออนโทโลยีนั้นสามารถกำหนดโครงสร้างข้อมูลในลักษณะลำดับชั้น (Hierarchical) และสามารถอธิบายข้อมูล (Metadata) ที่มีความสัมพันธ์ในฐานความรู้ได้ ลักษณะโครงสร้างการอธิบายจะอยู่ในรูปของคลาส คุณสมบัติของคลาส และความสัมพันธ์ของคลาส เพื่ออธิบาย Entity และความสัมพันธ์ (Relationship) ต่าง ๆ ที่เกิดขึ้น ซึ่งข้อดีคือทำให้คอมพิวเตอร์สามารถเข้าใจความหมายข้อมูลร่วมกันได้ ในงานวิจัยนี้กำหนดขอบเขตข้อมูลและเอกสารต่าง ๆ สำหรับใช้ทดสอบคือเอกสารงานวิจัยทางวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ซึ่งจะเก็บข้อมูลอยู่ในฐานข้อมูล (Database) และได้ออกแบบออนโทโลยีตามขอบเขตที่กำหนดไว้ โดยมีความสัมพันธ์ระหว่างคลาสเป็นแบบ a/0 ทั้งหมด ซึ่งตัวอย่างออนโทโลยีที่ได้ออกแบบไว้ในงานวิจัยบางส่วนแสดงดังภาพที่ 4 ซึ่งในส่วนการเพิ่มประสิทธิภาพการค้นคืนสารสนเทศเชิงความหมายนอกจาก Ontology-base ที่ได้ทำการออกแบบไว้ ยังมีส่วนที่ค่อยช่วยเพิ่มประสิทธิภาพและแปลงค่าข้อมูลอีก ซึ่งมีรายละเอียดดังนี้

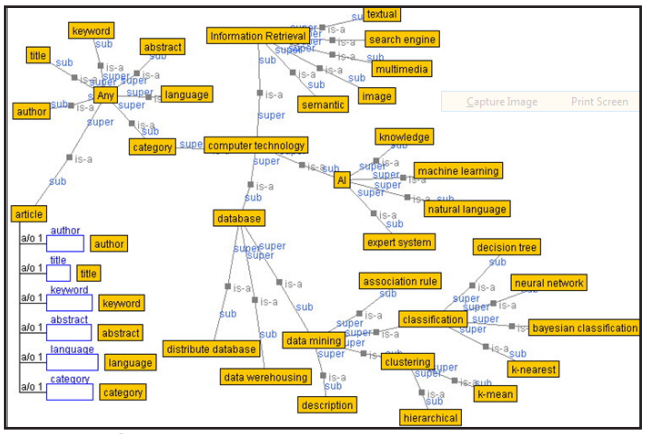
3.2.1 ฐานข้อมูล (Database) เป็นที่เก็บรวบรวมข้อมูลเอกสารเอกสารงานวิจัยทางวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ

3.2.2 Wrapper จะทำหน้าที่แปลงค่าจากฐานข้อมูล (Database) ให้อยู่ในออนโทโลยีที่นิยามด้วยภาษา OWL ซึ่งมีโครงสร้างตามใน Ontology-base ที่ได้ออกแบบไว้

3.2.3 Local Ontology คือ ออนโทโลยีที่มีโครงสร้างเหมือนกับใน Ontology-base ที่ได้ออกแบบไว้ ซึ่งเกิดจากการแปลงค่าจากข้อมูลในฐานข้อมูลมาเป็นออนโทโลยี

3.2.4 Ontology Mapping Module เป็นกระบวนการในการเชื่อมโยง Ontology-base และ Local Ontology เข้าด้วยกัน เพื่อช่วยให้สามารถค้นคืนข้อมูลได้

ในงานวิจัยนี้ได้ใช้ภาษา SPARQL สำหรับการค้นคืนข้อมูล ซึ่งภาษา SPARQL เป็นภาษาที่ใช้สำหรับสืบค้นข้อมูลจากออนโทโลยีเชิงความหมาย ซึ่งจะมีการเข้าถึงข้อมูลโดยอาศัยโครงสร้าง Triple (Subject, Predicate, Object) ซึ่งข้อดีของการใช้ SPARQL นั้น จะทำให้สามารถค้นหาข้อมูลโดยดูจากความสัมพันธ์ของข้อมูลต่าง ๆ ได้ซึ่งทำให้สามารถค้นหาเอกสารที่เกี่ยวข้องกับคิวรี่ได้ถึงแม้จะไม่มีในเอกสารจะไม่มีคำที่ค้นหานั้นปรากฏอยู่ก็ตาม



***ภาพที่ 4** ตัวอย่างส่วนหนึ่งของออนโทโลยีงานวิจัย*

3.3 การเรียงลำดับผลลัพธ์สำหรับระบบค้นคืนสารสนเทศข้ามภาษา (CLIR Result Ranking)

ในส่วนนี้ เมื่อทำการค้นคืนข้อมูลในออนโทโลยีเสร็จแล้ว ระบบจะทำการตรวจสอบผลลัพธ์ที่ได้จากการค้นคืนว่าระบบพบข้อมูลที่ผู้ใช้ต้องการหรือไม่ หลังจากนั้นระบบจะทำการจัดเรียบลำดับผลลัพธ์ ในงานวิจัยนี้ได้ปรับปรุงวิธีการคำนวณเพื่อให้การเรียงลำดับข้อมูลสำหรับเอกสารต่างภาษากันมีประสิทธิภาพมากยิ่งขึ้น แนวคิดสำคัญคือการให้น้ำหนักกับองค์ประกอบ 3 ส่วนที่แตกต่างกัน เช่น น้ำหนักคำสำคัญในเอกสาร (d) น้ำหนักคำสำคัญในคิวรี่ (q) และน้ำหนักของภาษา (l) การให้น้ำหนักของภาษาจะมีค่าเป็น 1 เมื่อคำสำคัญ

และเอกสารถูกเขียนด้วยภาษาเดียวกัน และเป็น 0.5 เมื่อคำสำคัญและเอกสารถูกเขียนด้วยคนละภาษา นอกจากนี้น้ำหนักของคำสำคัญในลำดับชั้นที่ต่างกันของออนโทโลยีที่ต่างกันจะมีค่าต่างกันด้วยเช่นกัน จากการทดลองพบว่าแนวคิดนี้สามารถเพิ่มประสิทธิภาพการจับลำดับให้มีประสิทธิภาพมากยิ่งขึ้น งานวิจัยนี้ปรับปรุงแนวคิดมาจากวิธี "ความคล้ายคลึงแบบโคซายน์ (Cosine Similarity)" ซึ่งเป็นวิธีการหาความคล้ายคลึงกันจากค่าความต่างของมุมของข้อมูล 2 อันที่เกิดขึ้นบนพื้นที่เว็กเตอร์ (Vector Space) ข้อดีของวิธีนี้คือจะทำให้เกิดความเป็นธรรม (fair) กับเอกสารที่มีความยาวไม่เท่ากัน วิธีการนี้เป็นที่นิยมและมีประสิทธิภาพสูงในการวัดความคล้ายคลึงระหว่างข้อมูล 2 อันและถูกนำมาประยุกต์ใช้กับศาสตร์การค้นคืนสารสนเทศอย่างแพร่หลาย [15] นักวิจัยได้ปรับปรุงสมการโคซายน์ โดยเพิ่มการนำค่าน้ำหนักของภาษา (l) ในคิวรี่มาคำนวณด้วย ดังสมการที่ (1)

$$sim(d, q) = \frac{\left(\sum_{i=1}^{n} d_i \times q\right)}{\left(\sum_{i=1}^{n} d_i^2\right)^{1/2} \times \left(\sum_{i=1}^{n} q^2\right)^{1/2}} \text{ x } 1 \qquad (1)$$

โดยที่ d คือ ค่าน้ำหนักของโหนดที่ค้นหาพบในออนโทโลยี
q คือ ค่าน้ำหนักของคำสำคัญที่ใช้ค้นหา
l คือ ค่าน้ำหนักของภาษา

การทำงานของอัลกอริทึม 2 อธิบายโดยสรุปได้ดังนี้ ขั้นตอนการเรียงลำดับนั้น จะทำการเรียงลำดับผลลัพธ์ที่ได้จากการค้นหา โดยใช้ค่าน้ำหนักของคำสำคัญ ค่าน้ำหนักของโหนดที่เจอในออนโทโลยี และค่าน้ำหนักของภาษามาคำนวณตามสมการที่ (1) แล้วจัดเรียงเก็บไว้ในอาเรย์

```
1 : keyword = array(keyword,weight)
2 : result = array(results from search, weight of node in Ontology,
3 : value of language)
4 : for each result
5 : score = calculate  Cosine Similarity from (1)
6 : result_ranking = array(results from search,score)
7 : sort result_ranking from score
8 : end for each
```

อัลกอริทึม 2 ขั้นตอนการเรียงลำดับในระบบค้นคืนข้ามภาษาที่ได้พัฒนาขึ้น ขั้นตอนสุดท้ายระบบจะส่งผลลัพธ์ที่ได้ทำการจัดลำดับไว้แล้วกลับมาแสดงให้กับผู้ใช้

4. การทดลองและผลการทดลอง

การทดลองสำหรับงานวิจัยนี้จะจำกัดทดลองกับข้อมูลเอกสารงานวิจัยทางด้านคอมพิวเตอร์ แต่เนื่องจากยังไม่มีองค์กรใดๆ ในประเทศไทยเก็บรวบรวมฐานข้อมูลเพื่อทดสอบระบบค้นคืนสารสนเทศข้ามภาษาที่เป็นมาตรฐาน ดังนั้นผู้วิจัยจึงทำการเก็บรวบรวมข้อมูลขึ้นเองจากแหล่งข้อมูลดังต่อไปนี้ www.ieee.org www.scopus.com และ www.thailis.or.th ในขั้นต้นของการทดลองมีจำนวนเอกสารงานวิจัยในฐานข้อมูลจำนวน 300 งานวิจัย แบ่งเป็นงานวิจัยที่ภาษาไทย 100 งานวิจัย และงานวิจัยภาษาอังกฤษ 200 งานวิจัย ในการทดลองนี้จะทำการทดสอบประสิทธิภาพของระบบ 2 ด้านเท่านั้น คือ ประสิทธิภาพการค้นหาและการจัดลำดับข้อมูลผลลัพธ์ เนื่องจากเป็นส่วนสำคัญสำคัญที่สุดของระบบค้นคืนสารสนเทศ

4.1 ประสิทธิภาพการค้นหาข้อมูล

ประสิทธิภาพสำคัญของระบบค้นคืนสารสนเทศ คือ ระบบสามารถค้นหาข้อมูลที่ผู้ใช้ต้องการได้อย่างถูกต้อง โดยค่ามาตรฐานที่นิยมใช้วัดประสิทธิภาพการค้นหาของระบบค้นคืนสารสนเทศคือ Precision และ Recall แสดงดังสมการที่ (2) และ (3) ตามลำดับ

$$Precission = \frac{\text{จำนวนเอกสารที่เกี่ยวข้องและถูกดึงออกมา}}{\text{จำนวนเอกสารที่ถูกดึงออกมาทั้งหมด}} \qquad (2)$$

$$Precission = \frac{\text{จำนวนเอกสารที่เกี่ยวข้องทั้งหมดและถูกดึงออกมา}}{\text{จำนวนเอกสารที่เกี่ยวข้องทั้งหมด}} \qquad (3)$$

ผู้วิจัยได้ทำการออกแบบการทดลองโดยใช้คำสำคัญที่แตกต่างกันจำนวนคำสำคัญ 20 คำ โดยแบ่งเป็นคำสำคัญภาษาไทย 10 คำ และคำสำคัญภาษาอังกฤษ 10 คำ โดยจะแบ่งการประเมินด้านออกค้นหาออกเป็น 6 ด้าน คือ

4.1.1 ด้านการค้นหาด้วยคำสำคัญภาษาไทย ในงานวิจัยที่เขียนด้วยภาษาไทย (QrTT)

4.1.2 ด้านการค้นหาด้วยคำสำคัญภาษาอังกฤษ ในงานวิจัยที่เขียนด้วยภาษาอังกฤษ (QrEE)

4.1.3 ด้านการค้นหาด้วยคำสำคัญภาษาไทย ในงานวิจัยที่เขียนด้วยภาษาอังกฤษ (QrTE)

4.1.4 ด้านการค้นหาด้วยคำสำคัญภาษาอังกฤษ ในงานวิจัยที่เขียนด้วยภาษาไทย (QrET)

4.1.5 ด้านการค้นหาด้วยคำสำคัญภาษาไทย ในงานวิจัยทั้งหมด (QrT)

4.1.6 ด้านการค้นหาด้วยคำสำคัญภาษาอังกฤษ ในงานวิจัยทั้งหมด (QrE)

ตัวอย่างระบบค้นคืนข้อมูลแสดงดังภาพที่ 5 แสดงภาพตัวอย่างจากการค้นหาโดยใช้ระบบค้นคืนข้ามภาษาที่ได้พัฒนาขึ้นโดยทดลองใส่คำสำคัญ “ออนโทโลยี” โดยระบบจะคำนวณค่า Precision และ Recall โดยจำนวนเอกสารที่เกี่ยวข้องและถูกดึงออกมานั้นจะดูจากผลลัพธ์ที่ระบบดึงออกมาแล้วตัดผลลัพธ์ที่ไม่ถูกออกโดยผู้วิจัยอีกที และจำนวนเอกสารที่เกี่ยวข้องทั้งหมดนั้นได้เช็คและนับจากฐานข้อมูลโดยผู้วิจัยเอง โดยเอกสารที่จัดอยู่ในลำดับแรกนั้นเป็นเอกสารที่มีความสัมพันธ์มากที่สุด เนื่องจากมีคำสำคัญปรากฏอยู่ในทุกส่วน ทั้งส่วนของชื่องานวิจัย คำสำคัญ และในบทคัดย่อ ซึ่งเมื่อคำนวณตามสมการที่ (1) แล้วทำให้มีคะแนนในการจัดลำดับอยู่ในลำดับที่สูงสุด และในลำดับต่อมาก็จะมีคะแนนการจัดลำดับลดลั่นกันไป สำหรับการเรียงลำดับข้อมูลนั้นจะอธิบายในหัวข้อถัดไป ผลการทดลองนี้แสดงดังตารางที่ 2

ตารางที่ 2 *ผลการทดลอง*

	QrTT	QrEE	QrTE	QrET	QrT	QrE
Recall	0.92	0.95	0.88	0.82	0.91	0.90
Precision	0.92	0.96	0.86	0.82	0.88	0.91

จากผลทดลองในตารางที่ 2 นั้น จะเห็นได้ว่า ค่า Recall และ Precision มากที่สุดอยู่ที่ด้านการค้นหาด้วยคำสำคัญภาษาอังกฤษในงานวิจัยที่เขียนด้วยภาษาอังกฤษ ซึ่งมีค่าคือ 0.95 และ 0.96 ตามลำดับซึ่งถือว่ามีประสิทธิภาพในการค้นหาที่ดี เพราะว่าจากการขยายคำสำคัญคำภาษาอังกฤษนั้นทำให้ได้จำนวนคำสำคัญที่ใช้ค้นหาเพิ่มขึ้นและครอบคลุมในความหมายเดียวกัน ทำให้สามารถค้นหาเอกสารที่มีเนื้อหาเกี่ยวข้องได้อย่างครอบคลุม ส่วนค่าที่น้อยที่สุดอยู่ที่ด้านการค้นหาด้วยคำสำคัญภาษาอังกฤษในงานวิจัยที่เขียนด้วยภาษาไทย ซึ่งมีค่า Recall คือ 0.82 และค่า Precision คือ 0.82 ตามลำดับ สาเหตุที่ด้านนี้ได้ค่าน้อยที่สุด เนื่องจากในการขยายคิวรี่นั้น ThaiWordnet ที่นำมาใช้นั้นอยู่ในช่วงของการพัฒนาไปด้วย ทำให้อาจจะไม่ครอบคลุมคำศัพท์เฉพาะบางคำได้ ทำให้ไม่สามารถค้นหาเอกสารที่เกี่ยวข้องได้ทั้งหมด

จะสังเกตได้ว่าจากการค้นหาในภาพที่ 5 นั้น จะเห็นได้ว่าผลลัพธ์ในลำดับที่ 4 นั้นไม่ได้มีคำสำคัญปรากฏอยู่ แต่ก็สามารถจะค้นหาพบ เพราะว่าการใช้ออนโทโลยีนั้นทำให้สามารถค้นหาเอกสารที่เกี่ยวข้องกันเชิงความหมายได้ แม้จะไม่มีคำสำคัญปรากฏอยู่ก็ตาม เช่นดังในตัวอย่างออนโทโลยี และ Semantic มีความสัมพันธ์กันเชิงความหมายตามที่ได้ออกแบบไว้ในออนโทโลยี ระบบจึงสามารถค้นคืนเอกสารออกมา

4.2 ประสิทธิภาพการเรียงลำดับข้อมูล

ในการทดลองด้านการจัดเรียงผลลัพธ์นั้น ได้ใช้สมการ (1) เข้ามาช่วยในการคำนวณค่าความคล้ายคลึง จากภาพที่ 5 ซึ่งแสดงผลการค้นหาออกมานั้น สามารถคำนวณค่าความคล้ายคลึงกันซึ่งข้อมูลที่คำนวณออกมานั้นดังแสดงในตารางที่ 3 ซึ่งเป็นข้อมูลที่ได้จากการค้นหา เมื่อทำการขยายคิวรี่จะได้ q = ออนโทโลยี, ontology และสามารถคำนวณออกมาได้ดังนี้

ตารางที่ 3 *ข้อมูลที่ได้จากการค้นหา*

คำ / เอกสาร	doc 1 (d1)	doc 2 (d2)	doc 3 (d3)	doc 4 (d4)	q
ออนโทโลยี	0.8	0.8	0	0	1
ontology	0.8	0.8	0.8	0.8	1
Cross language	0	0	1	1	0
ข่าวออนไลน์	0.8	0	0	0	0
Text Analysis	1	0	0	0	0
เวิร์ดเน็ต	0	1	0	0	0
dictionary	0	0	0	1	0
การวิเคราะห์ข้อความ	1	0	0	0	0

จากข้อมูลที่ได้จากการค้นหาในตารางที่ 3 จะคำนวณเฉพาะที่ค่า q ไม่เท่ากับ 0 สามารถคำนวณได้ดังต่อไปนี้

$$sim(q, doc1, l) = \frac{(0.8x1)+(0.8x1)}{((0.8^2+0.8^2)x(1^2+1^2))^{\frac{1}{2}}} \text{ x } 1 = 1 \quad (4)$$

$$sim(q, doc2, l) = \frac{(0.8x1)+(0.8x1)}{((0.8^2+0.8^2)x(1^2+1^2))^{\frac{1}{2}}} \text{ x } 1 = 1 \quad (5)$$

$$sim(q, doc3, l) = \frac{0.8x1}{(0.8^2 \text{ x } 1^2)^{\frac{1}{2}}} \text{ x } 0.5 = 0.5 \quad (6)$$

$$sim(q, doc4, l) = \frac{0.8x1}{(0.8^2 \text{ x } 1^2)^{\frac{1}{2}}} \text{ x } 0.5 = 0.5 \qquad (7)$$

ซึ่งจากการคำนวณค่าความคล้ายคลึงนั้น จะเห็นได้ว่าสามารถเรียงลำดับผลลัพธ์ได้ตามความสัมพันธ์จากคำสำคัญที่ใช้ โดยค่าน้ำหนักของโหนดที่พบและค่าน้ำหนักของคำสำคัญนั้น จะมีผลต่อการเรียงลำดับมากที่สุด ซึ่งสำหรับค่าน้ำหนักของคำสำคัญ (d) นั้น เมื่อได้ขยายคิวรี่แล้ว คำที่นำมาค้นหานั้น ก็จะมีค่าน้ำหนักที่ต่างกันไปตามที่ได้ให้น้ำหนักไว้ในขั้นตอนการขยายคิวรี่ ซึ่งหากไม่นำน้ำหนักของภาษามาคำนวณด้วย จะมาสามารถคำนวณได้ดังนี้

$$sim(q, doc1) = \frac{(0.8x1)+(0.8x1)}{((0.8^2+0.8^2)x(1^2+1^2))^{\frac{1}{2}}} = 1 \qquad (8)$$

$$sim(q, doc2) = \frac{(0.8x1)+(0.8x1)}{((0.8^2+0.8^2)x(1^2+1^2))^{\frac{1}{2}}} = 1 \qquad (9)$$

$$sim(q, doc3) = \frac{0.8x1}{(0.8^2 \text{ x } 1^2)^{\frac{1}{2}}} = 1 \qquad (10)$$

$$sim(q, doc4) = \frac{0.8x1}{(0.8^2 \text{ x } 1^2)^{\frac{1}{2}}} = 1 \qquad (11)$$

จะเห็นได้ว่าเมื่อคำนวณโดยไม่คิดค่าน้ำหนักนั้น คะแนนที่ได้จากการคำนวณนั้นจะมีค่าเท่ากันทั้งสองภาษา ซึ่งอาจจะทำให้เอกสารที่ถูกเขียนด้วยภาษาเดียวกับคำสำคัญนั้นไม่ได้ถูกจัดลำดับไว้ในลำดับแรก

5. สรุปงานวิจัย

งานวิจัยนี้นำเสนอแนวคิดในการพัฒนาระบบค้นหาข้ามภาษาระหว่างภาษาไทยและภาษาอังกฤษ โดยใช้เทคนิคขยายคิวรี่ (Query Expansion) เข้ามาช่วยซึ่งสามารถขยายคำสำคัญที่ใช้ค้นหาได้มากขึ้น ทำให้มีคำสำคัญเพิ่มขึ้นสำหรับการค้นหา และได้ใช้เทคนิคออนโทโลยีเข้ามาช่วยทำให้สามารถค้นหาข้อมูลเชิงความหมายได้ และแนวคิดที่สองคือ การนำเสนอวิธีการจัดลำดับข้อมูลสำหรับระบบค้นคืนข้อมูลข้ามภาษาโดยนำข้อมูลลำดับชั้นของออนโทโลยีภาษาที่ใช้ และ synonym มาพิจารณาการคำนวณเพื่อจัดลำดับทำให้การจัดลำดับผลลัพธ์มีประสิทธิภาพมากยิ่งขึ้น

อย่างไรก็ตามงานวิจัยนี้ยังมีข้อจำกัดในส่วนของรองรับความไม่สมบูรณ์ของออนโทโลยี (Ontology Imperfection) เนื่องจากการสร้างออนโทโลยีไม่สามารถสร้างให้ครอบคลุมองค์ความรู้ทุกอย่างในโดเมนได้ในครั้งแรก และถึงแม้จะมีการแก้ไขออนโทโลยีหลายๆ ครั้งก็ยังเป็นไปได้ยากที่จะครอบคลุมองค์ความรู้ทั้งหมด [16] และมีต้นทุนด้านแรงงานที่สูงมาก ดังนั้นระบบค้นคืนสารสนเทศที่เก็บข้อมูลในออนโทโลยีควรคำนึงถึงปัญหานี้และมีเครื่องมือรองรับเมื่อไม่สามารถหาข้อมูลที่ผู้ใช้ต้องการภายในออนโทโลยีได้ ซึ่งผู้วิจัยจะทำการพัฒนาเทคนิคเพื่อรองรับปัญหานี้ต่อไปในอนาคต

6. เอกสารอ้างอิง

[1] ทัศนวรรณ ศูนย์กลาง และคณะ. “การเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษ เพื่อการค้นคืนข้ามภาษาด้วยเทคนิคนิวรอลเน็ตเวิร์ก.” *งานประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ ครั้งที่ 4 (NCSEC 2000),* 2000.

[2] A. Chen. “Multilingual Information Retrieval Using English and Chinese Queries Evaluation of Cross-Language Information Retrieval Systems.” *Evaluation of Cross-Language Information Retrieval Systems*, Vol. 2406, pp. 373-381, 2002.

[3] R. F. E. Sutcliffe, I. Gabbay and A. O’Gorman. “Cross-Language French-English question answering using the DLT system at CLEF 2003.” *Comparative Evaluation of Multilingual Information Access Systems*, Vol. 3237, pp. 572-580, 2004.

[4] R. F. E. Sutcliffe, I. Gabbay, M. Mulcahy and A.O’Gorman. “Cross-language French-English Question Answering using the DLT system at CLEF 2004.” *In Proceedings of the 5th conference on Cross-Language Evaluation Forum: multilingual Information Access for Text*, *Speech and Images,* pp. 404-410, 2005.

[5] D. Zhou and V. Wade. “The Effectiveness of Results Re-Ranking and Query Expansion in Cross-language Information Retrieval.” *In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering*

and Cross-Lingual Information Access, 2010.

[6] E. Celebi, B. Sen and B. Gunel. "Turkish - English cross language information retrieval using LSI." *In Proceedings of the 24th International Symposium on Computer and Information Sciences (ISCIS 2009),* pp. 634-638, 2009.

[7] D. Zhou, S. Lawless, J. Min and V. Wade. "A late fusion approach to cross-lingual document re-ranking." *In Proceedings of the 19th ACM international conference on Information and knowledge management,* pp. 1433-1436, 2010.

[8] D. W. Embley, S. W. Liddle, D. W. Lonsdale and Y. Tijerino. "Multilingual ontologies for cross-language information extraction and semantic search." *Conceptual Modeling – ER 2011,* Vol. 6998 LNCS, pp. 147-160, 2011.

[9] W. Gao, C. Niu, M. Zhou and K. F. Wong. "Joint ranking for multilingual web search." *In Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pp. 114-125, 2009.

[10] T. Y. Liu. "Learning to rank for information retrieval." *Foundations and Trends in Information Retrieval*, Vol. 3, pp. 225-331, 2009.

[11] D. Zhou and V. Wade. "Latent Document Re-Ranking." *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Vol. 3, pp. 1571-1580, 2009.

[12] Y. L. Cao, T. J. Huang and Y. H.Tian. "A ranking SVM based fusion model for cross-media meta-search engine." *Journal of Zhejiang University: Science C,* Vol. 11, pp. 903-910, 2010.

[13] Y. Cao, J. Xu, T. Y. Liu, H. Li, Y. Huang and H. W. Hon. "Adapting Ranking SVM to Document Retrieval." *In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 186-193, 2006.

[14] S. Thoongsup, K. Robkop, C. Mokarat, T. Sinthurahat, T. Charoenporn, V. Sornlertlamvanich and H. Isahara. "Thai WordNet construction." *In Proceedings of the 7th Workshop on Asian Language Resources*, 2009.

[15] ไกรศักดิ์ เกษร. "การค้นหาข้อมูลเชิงความหมาย: แนวคิดใหม่ของโปรแกรมการค้นหา (Search Engine) และแนวทางการพัฒนาในอนาคต." *วารสารวไลยอลงกรณ์ปริทัศน์*, Vol. 2, 2011.

[16] G. Nagypál. "Improving information retrieval effectiveness by using domain knowledge stored in ontologies." *In Proceedings of the 2005 OTM Confederated international conference,* pp. 780-789, 2005.