# การใช้กราฟความหมายสำหรับสกัดคำสำคัญจากเอกสาร ในเครื่องมือค้นหาข้อมูลเชิงลึก

รัชฎา คงคะจันทร์[1*] วสิศ ลิ้มประเสริฐ[1] ปกป้อง ส่องเมือง[1] ชัยณรงค์ เกษามูล[2]

[1]สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี
มหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต ปทุมธานี
[2]สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี
มหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต ปทุมธานี
[*]Corresponding author email: krachada@tu.ac.th

ได้รับบทความ: 27 พฤศจิกายน 2566
ได้รับบทความแก้ไข: 11 กุมภาพันธ์ 2567
ยอมรับตีพิมพ์: 29 มีนาคม 2567

**บทคัดย่อ**

บทความนี้ นำเสนอการใช้กราฟความหมายเป็นแหล่งความรู้ในการเปรียบเทียบเชิงความหมายระหว่างคำสืบค้นและเนื้อหาของเอกสาร กราฟความหมายถูกสร้างโดยอัตโนมัติโดยใช้หลักการทำงานทางด้านการประมวลผลภาษาธรรมชาติ โดยขั้นตอนในการสร้างกราฟความหมายเริ่มจาก การเตรียมเอกสารเบื้องต้น ซึ่งได้แก่ การตัดคำ การลบคำหยุด และกำกับหน้าที่ของคำ จากนั้นเป็นขั้นตอนการแจกแจงประโยคโดยใช้วิธีการแจกแจงไวยากรณ์แบบพึ่งพา เพื่อหาความสัมพันธ์ของคำที่อยู่ในประโยค ต่อมาเป็นการแปลงความสัมพันธ์ระหว่างคำที่ได้เป็นกราฟความหมายซึ่งอยู่ในรูปของกราฟมโนทัศน์ และในขั้นตอนสุดท้าย กราฟความหมายที่ได้จะถูกนำมาคำนวณเพื่อหาค่าความสัมพันธ์ของคำและประโยคในเอกสาร เพื่อสกัดเอาคำสำคัญออกมา โดยงานวิจัยนี้ได้ทดลองสร้างกราฟความหมายจากเอกสาร จำนวน 380 เอกสาร ที่รวบรวมมาจาก บทความวิจัยในเว็บไซต์ของ IEEE และ จำนวน 144 เอกสาร จากคลังข้อมูลมาตรฐาน SemEval โดยผลที่ได้จากการสกัดคำสำคัญแบบอัตโนมัติ มีค่าความแม่นยำ ร้อยละ 40 ค่าความครบถ้วน ร้อยละ 53 และ F1-Measure ที่ ร้อยละ 40 ตามลำดับ

**คำสำคัญ:** การค้นหาแนวตั้ง/ เครื่องมือค้นหาโดยใช้ความหมาย/ กราฟมโนทัศน์/ การสกัดคำสำคัญ/ กราฟความหมายของเอกสาร

# Using Semantic Graph for Keyword Extraction in Vertical Search Engine

Rachada Kongkachandra[1*], Wasit Limprasert[1], Pokpong Songmuang[1], Chainarong Kesamoon[2]

[1] Department of Computer Science, Faculty of Science and Technology, Thammasat University Rangsit Centre, Pathum Thani
[2]Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University Rangsit Centre, Pathum Thani
[*]Corresponding author email: krachada@tu.ac.th

Received: 27 November 2023
Revised: 11 February 2024
Accepted: 29 March 2024

**Abstract**

This paper presents the semantic graphs as knowledge resource for keyword extraction. The semantic graph generation process starts from the preprocessing steps i.e. word tokenization, stop word removing and part-of-speech tagging. Then the sentences are parsed by using dependency parsing technique. Next, the parsed sentences are represented into semantic graph by considering their word relations. In this paper, the semantic graphs are represented using the conceptual graph principle. Finally, the semantic graphs are scored based on their semantic relatedness. The words in the high score semantic graphs are extracted as the keywords. The semantic graphs and keywords are generated and extracted using 380 documents from the IEEE website and 144 documents from SemEval standard corpus. The precision, recall and F1 scores are 40%, 53%, and 40%, respectively.

**Keywords:** Vertical Search/ Semantic Search Engine/ Conceptual Graph/ Keyword Extraction/ Semantic Graph

**บทนำ**

ยุคปัจจุบันนี้เป็นยุคของข้อมูลขนาดใหญ่ ที่ข้อมูลดิจิทัลเกิดขึ้นอย่างมากมายหลากหลาย และ รวดเร็ว ทำให้การสืบค้นข้อมูลด้วยวิธีเดิมไม่สามารถให้คำตอบได้ในเวลาที่รวดเร็ว ด้วยเหตุนี้เครื่องมือค้นหาข้อมูล (Search Engine) จึงมีความจำเป็นอย่างมาก ซึ่งในปัจจุบันมีเครื่องมือค้นหาข้อมูลให้ใช้งานอยู่หลายแบรนด์ อาทิเช่น กูเกิ้ล [1] และ บิงของไมโครซอฟต์ [2] ซึ่งเครื่องมือค้นหาข้อมูลเหล่านี้ได้พยายามปรับปรุง อัลกอริทึมและเพิ่มฟังก์ชันการทำงาน เพื่อตอบสนองความสะดวกสบายให้กับผู้ใช้มากขึ้น หลักการทำงานของเครื่องมือค้นหาข้อมูลโดยรวม คือจะนำเอาคำสืบค้น (Query) ที่ผู้ใช้ป้อน ไปเปรียบเทียบกับเนื้อหาในเอกสารที่อยู่ในอินเตอร์เน็ต จากนั้นนำเอาเอกสารมาเรียงลำดับตามความเกี่ยวข้องกับคำที่สืบค้นจากมากไปน้อย แต่เนื่องจากส่วนใหญ่แล้วเอกสารจะประกอบไปด้วยคำจำนวนมาก ซึ่งถ้าหากใช้คำทุกคำมาเปรียบเทียบจะทำให้เสียเวลาในการประมวลผล ดังนั้น เครื่องมือค้นหาข้อมูลส่วนใหญ่จะแปลงเอกสารให้อยู่ในรูปแบบที่เล็กลงแต่ยังคงความหมายของเอกสารเช่นเดิม แล้วใช้เป็นตัวแทนของเอกสารนั้น ซึ่ง คำสำคัญ (Keyword) มักจะถูกนำมาใช้เป็นตัวแทนเอกสาร ดังปรากฏเป็นตัวอย่างใน [3] [4] และ [5]

การสกัดคำสำคัญจากเอกสาร (Keyword Extraction) เป็นการประยุกต์เอาศาสตร์ทางปัญญาประดิษฐ์มาใช้ในการสกัดคำสำคัญจากเอกสารโดยอัตโนมัติ ซึ่งความยากของการดำเนินการคือ จะทำอย่างไรให้คอมพิวเตอร์สามารถรู้ได้ถึงความแตกต่างระหว่าง คำสำคัญ กับ คำทั่วไป ในเอกสาร ที่ผ่านมามีงานวิจัยหลายงานได้ เสนอเทคนิคในการสกัดคำสำคัญโดยอัตโนมัติ ซึ่งบทความนี้ได้แบ่งออกเป็นกลุ่มใหญ่ ๆ ได้ 3 กลุ่มด้วยกันตามเกณฑ์ของข้อมูลที่ใช้ในการพิจารณา คือ (1) กลุ่มพิจารณาโดยใช้กฎไวยากรณ์ของภาษา (Linguistic Rules) โดยรูปแบบของคำสำคัญจะถูกกำหนดโดยคุณลักษณะทางภาษา เช่น ในงานของ [6] ซึ่งได้เลือกคำหรือกลุ่มคำที่ทำหน้าที่เป็นคำนาม และ คำกริยา ของเอกสารเป็นคำสำคัญ ซึ่งจะใช้ผู้เชี่ยวชาญทางด้านภาษาศาสตร์เป็นผู้กำหนดกฎเหล่านั้น (2) กลุ่มพิจารณาโดยใช้คลังข้อมูลฝึกฝน (Training Corpus) [7-9] สำหรับกลุ่มนี้ คำสำคัญจะถูกพิจารณาโดยดูจากคุณลักษณะเชิงสถิติ เช่น จำนวนครั้งที่เกิดของคำภายในและระหว่างเอกสาร (Term Frequency and Inverse Document Frequency) หรือ จำนวนครั้งของคำที่เกิดร่วมกัน (Number of Co-occurrence) ซึ่งอาจจะใช้เป็นข้อมูลเชิงสถิติอย่างเดียว [8] หรือใช้หลักของการเรียนรู้ของเครื่องมาร่วมด้วย [7] และ [9] โดยจะใช้ข้อมูลจากคลังข้อมูลขนาดใหญ่ (Corpora) และ (3) กลุ่มพิจารณาโดยใช้ความหมาย (Semantics) [10-12] สำหรับกลุ่มนี้ คำสำคัญ จะถูกเลือกโดยพิจารณาที่ความสัมพันธ์ทางความหมาย (Word Relatedness) ระหว่างคำสำคัญกับเนื้อหาของเอกสาร ซึ่งเริ่มจากการพิจารณาความคล้ายกันของคำ (Word Similarity) ซึ่งส่วนใหญ่จะใช้ ฐานข้อมูลคำศัพท์ทางความหมายที่ชื่อว่า WordNet

ซึ่งพัฒนาโดย จอร์จ เอ มิลเลอร์ จาก มหาวิทยาลัย พรินซ์ตัน ในปี ค.ศ. 1995 เป็นแหล่งข้อมูลสำหรับเปรียบเทียบเชิงความหมาย ซึ่ง WordNet นั้นจะแสดงในลักษณะของโครงข่ายของคำ โดยคำที่มีความเกี่ยวข้องกันจะเชื่อมโยงกัน ที่ผ่านมานักวิจัยใช้ระยะห่างระหว่างคู่คำที่อยู่บนโครงข่าย WordNet เป็นเกณฑ์ตัดสินความคล้ายกันของคำ โดยใช้หลักการที่ว่า คำที่อยู่ใกล้กันมากกว่า จะมีความหมายคล้ายกันมากกว่าคำที่อยู่ห่างกัน ซึ่งนักวิจัยที่ผ่านมาได้คิดค้น วิธีการในการวัดระยะห่างของคู่คำ ได้แก่ งานวิจัยของ [10-12]

ถึงแม้ว่าในปัจจุบันนี้ เครื่องมือสำหรับค้นหาข้อมูลอย่าง Google จะสามารถตอบโจทย์ผู้ใช้ได้เป็นอย่างดี แต่ผลลัพธ์ที่ได้จากการค้นหาจะเป็นข้อมูลในแนวกว้าง เหมาะสำหรับการสืบค้นโดยทั่วไป แต่นับตั้งแต่ต้นศตวรรษที่ 21 การใช้งานอินเตอร์เน็ตเริ่มนำมาใช้สำหรับธุรกิจมากขึ้น ซึ่งการค้นหาข้อมูลของผู้ใช้เริ่มให้ความสนใจในแนวลึกมากขึ้น เครื่องมือค้นหาข้อมูลจะต้องสามารถเข้าใจคำสืบค้น และสามารถหาคำตอบที่เกี่ยวข้อง จากนั้นใช้ข้อมูลที่ได้จากการค้นหามาประกอบเพื่อหาข้อมูลที่มีความหมายลึกลงไปจนกระทั่งตอบโจทย์ของผู้ค้นหา เครื่องมือค้นหาข้อมูลแบบนี้เรียกว่า เครื่องมือค้นหาข้อมูลเชิงลึก (Vertical Search) [13-14] ซึ่งเครื่องมือค้นหาข้อมูลจะสามารถค้นหาในแนวลึกได้นั้น จำเป็นจะต้องนำเอาความหมายมาใช้ในการพิจารณา ซึ่งจะเป็นความหมายตามบริบท ดังนั้นในบทความนี้ จะนำเสนอ เทคนิคที่จะทำให้เครื่องมือค้นหาข้อมูลเชิงลึก สามารถเข้าใจความหมายของคำสืบค้นและเนื้อหาเอกสารได้ โดยจะเสนอรูปแบบของความหมายในระดับประโยค เพื่อนำไปใช้ในการสกัดคำสำคัญเชิงความหมาย เพื่อให้ผู้ใช้ได้รับข้อมูลที่ตรงกับความต้องการมากยิ่งขึ้น ซึ่งในหัวข้อที่ 2 จะกล่าวถึงวิธีการในการสร้างกราฟความหมายแบบอัตโนมัติ ซึ่งกราฟความหมาย (Semantic Graph) นี้จะถูกใช้เป็นแหล่งความรู้ที่คอมพิวเตอร์ใช้ในการเปรียบเทียบความหมายของเอกสารกับคำที่ผู้ใช้ ใช้สำหรับสืบค้น จากนั้นในหัวข้อที่ 3 จะเป็นวิธีการคำนวณคะแนนความคล้ายทางความหมายระหว่าง คำที่สืบค้นกับเนื้อหาของเอกสาร สำหรับหัวข้อที่ 4 จะกล่าวถึงวิธีการและผลการทดสอบประสิทธิภาพ ของวิธีการที่นำเสนอ และปิดท้ายด้วยการสรุปและอภิปรายผลของการทดสอบที่ได้

## วัสดุและวิธีการ

### 1. กลุ่มของเอกสารที่ใช้ในการวิจัย

เอกสารที่ใช้สำหรับการทดลองนี้ใช้ข้อมูลจาก คลังข้อมูลมาตรฐาน SemEval [15] และ บทความจากเว็บไซต์ IEEE โดยใช้เอกสารจาก SemEval จำนวน 144 เอกสาร และจากเว็บไซต์ IEEE จำนวน 380 เอกสาร รายละเอียดของเอกสารที่ใช้ แสดงในตารางที่ 1

**ตารางที่ 1** รายละเอียดของเอกสารที่ใช้ในการทดลอง

| คลังข้อมูล | เอกสาร | | | | คำสำคัญ | | |
|---|---|---|---|---|---|---|---|
| | ประเภท | ภาษา | จำนวนเอกสาร | จำนวนคำ (เฉลี่ย) | จำนวนรวม | จำนวนเฉลี่ย | คำที่หายไป (%) |
| SemEval | บทคัดย่อ | อังกฤษ | 144 | 85.01 | 6,251 | 43.41 | 18.87 |
| IEEE | บทคัดย่อ | อังกฤษ | 380 | 72.87 | 14,607 | 38.44 | 42.00 |

**2. เครื่องมือที่ใช้ในงานวิจัย**

ส่วนของการเตรียมข้อมูล และ แจกแจงไวยากรณ์ของเอกสาร จะใช้ ไลบรารี่ของ Spacy (https://spacy.io/api/dependencyparser) ส่วนการทดลองในการสกัดคำสำคัญ จะใช้ ไลบรารี่ของ Pke (https://github.com/varun112jain/pke)

**3. การใช้กราฟมโนทัศน์ในการแทนความหมายของประโยค**

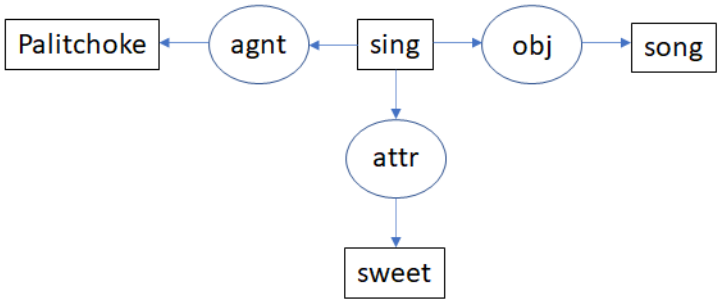
การค้นหาข้อมูลเชิงความหมาย (Semantic Search) เป็นหลักการค้นหาที่มักจะนำมาใช้ในเครื่องมือค้นหาข้อมูลเชิงลึก โดยจะใช้ความหมายมาเป็นคุณลักษณะในการเปรียบเทียบเพื่อหาคำตอบที่ต้องการ จะพยายามปรับปรุงความแม่นยำในการค้นหาด้วยการทำความเข้าใจเจตนาของผู้ค้นหาและความหมายตามบริบทของคำศัพท์ตามที่ปรากฏในปริภูมิสำหรับค้นหา (Search Space) เพื่อให้ได้ผลของการค้นหาที่เกี่ยวข้องมากที่สุด การค้นหาเชิงความหมายจะช่วยให้การค้นหาในแนวตั้ง เลือกข้อมูลที่ลึกและมีความถูกต้องมากขึ้น ตัวอย่างเช่น ถ้าเราอยู่ในเว็บไซต์อุปกรณ์คอมพิวเตอร์ แล้วใช้เครื่องมือค้นหาเชิงความหมาย หาคำว่า "mouse" เราจะได้ โฆษณาที่เกี่ยวข้องกับ Input device มากกว่าเป็นข้อมูลของ หนู ที่เป็นสัตว์

ในการนำเอาการค้นหาเชิงความหมายมาประยุกต์ใช้เพื่อพัฒนา เครื่องมือค้นหาแบบเชิงลึก นั้น จะนำไปใช้ในส่วนของ Indexing และ searching ของการพัฒนาเครื่องมือค้นหา ในส่วนของ Indexing นั้น เว็บไซต์ จะถูกจัดลำดับตามความเกี่ยวข้องกับ คำสืบค้นของผู้ใช้ ซึ่งการจัดอันดับของเว็บไซต์นั้น ถ้าเรียงตามความหมายของเนื้อหาของเว็บไซต์กับคำสืบค้น จะให้ผลลัพธ์ที่ตรงใจผู้ใช้ แต่เนื่องจาก เวลาที่ใช้ในการสืบค้นถือว่าเป็นปัจจัยที่สำคัญ ดังนั้น การจัดอันดับเว็บไซต์โดยดูจากเนื้อหาทั้งหมด จะใช้เวลามาก ดังนั้น จึงมีการนำเอา งานวิจัยด้าน การสกัดคำสำคัญ (Keyword Extraction) มาใช้เพื่อหาตัวแทนเอกสาร (เนื้อหาของเว็บไซต์) เพื่อลดเวลาในการสืบค้น และ ลดพื้นที่ในการจัดเก็บ

การจะทำให้คอมพิวเตอร์สามารถค้นหาข้อมูลได้ใกล้เคียงกับมนุษย์นั้น เราจำเป็นต้องสอนให้คอมพิวเตอร์เข้าใจความหมายเสียก่อน ดังนั้น หลักการของ การแทนองค์ความรู้ (Knowledge Representation) ซึ่งเป็นสาขาย่อยของปัญญาประดิษฐ์ จึงถูกนำมาประยุกต์ใช้ในโครงการวิจัยนี้ **กราฟมโนทัศน์ (Conceptual Graph)** [16] ซึ่งเป็นการแทนองค์ความรู้รูปแบบหนึ่งได้ถูกเลือกมาใช้สำหรับแทนความหมายในโครงการวิจัยนี้ เนื่องจากมีข้อดี คือเข้าใจได้ง่าย เพราะใช้กราฟในการนำเสนอ และสามารถอนุมานเหตุผลได้โดยอัตโนมัติ เพราะกราฟมโนทัศน์มีวิวัฒนาการมาจากตรรกศาสตร์แบบ Existential แนวคิดสำหรับบทความนี้ ข้อความหนึ่งประโยค จะถูกแทนด้วยกราฟมโนทัศน์ ซึ่งอาจจะเป็นหนึ่งกราฟหรือหลายกราฟ ขึ้นอยู่ความยาวของประโยค ยกตัวอย่างเช่น

**ผลิตโชคร้องเพลงไพเราะมาก**

**Palitchoke sings a song so sweet.**




**ภาพที่ 1** แสดงถึง กราฟความหมายของประโยค “ผลิตโชคร้องเพลงไพเราะมาก”

กราฟที่แสดง ในภาพที่ 1 มีโนดอยู่สองแบบ คือ โนด สี่เหลี่ยม และ โนด วงกลม โนดสี่เหลี่ยม เรียกว่า Concept Node สำหรับแทน วัตถุ แนวคิด หรือเหตุการณ์ โครงการวิจัยนี้เราใช้สำหรับแทนคำที่แสดงแนวคิด ซึ่งคำที่เลือกมาจะเป็น คำนาม คำกริยา คำวิเศษณ์ และ คำคุณศัพท์ ส่วน โนด วงกลมนั้น เรียกว่า Conceptual Relation Node สำหรับแทนความสัมพันธ์ระหว่าง แนวคิดสองแนวคิดที่แสดงโดยโนดสี่เหลี่ยม โดยความสัมพันธ์ระหว่างแนวคิด กำหนดโดยพิจารณาจากความสัมพันธ์ทางไวยากรณ์แบบพึ่งพา (Dependency Grammar) และปรับแต่งเพิ่มเติมโดยผู้เชี่ยวชาญ

**4. เทคนิคการสกัดคำสำคัญโดยใช้กราฟความหมาย**

การสกัดคำสำคัญของบทความนี้ ใช้หลักการของ Graph-based Keyword Extraction ที่มีคุณสมบัติพื้นฐานดังนี้

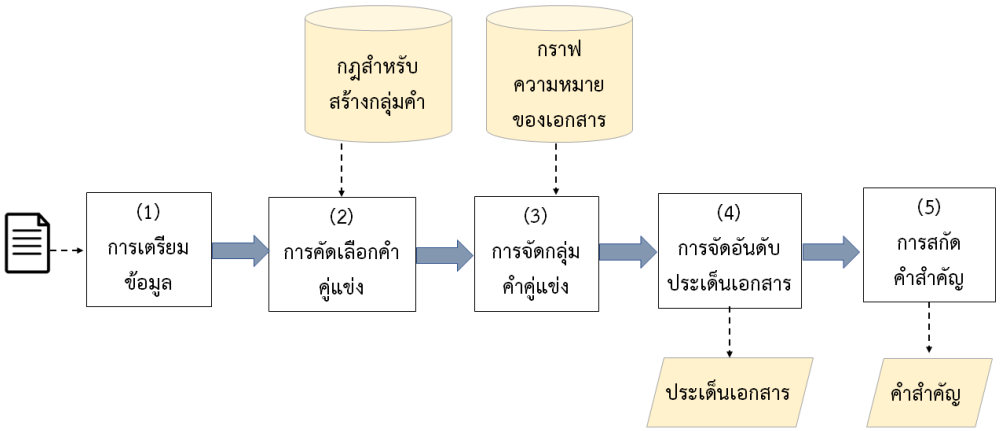
- เอกสารจะถูกนำเสนอในรูปแบบของ กราฟที่เชื่อมต่อกันแบบสมบูรณ์ เรียกว่า Complete Graph
- ข้อมูลที่อยู่ใน Node จะถูกนำมาพิจารณาและจัดอันดับ ตามความสำคัญ ซึ่งข้อมูลใน Node อาจจะเป็น คำ วลี หรือ ประเด็นของเอกสาร (Topic)
- ค่าน้ำหนักที่ Edge เป็นค่าความเกี่ยวข้องระหว่าง Node ที่เชื่อมกัน

สำหรับงานวิจัยนี้ ใช้หลักการเดียวกับ TopicRank [18] ที่กำหนดสมมุติฐานว่า ***"ความหมายของเอกสาร อาจประกอบไปด้วย ประเด็นเอกสาร (Topic) หนึ่งหรือหลายประเด็น และ คำสำคัญที่ดี ที่ทำหน้าที่เป็นตัวแทนของเอกสาร ควรจะครอบคลุมทุกประเด็นของเอกสาร"***

ดังนั้น โครงงานวิจัยนี้ ได้กำหนดรายละเอียดของ กราฟ ดังนี้

- ข้อมูลที่อยู่ในแต่ละ Node คือ ประเด็นของเอกสาร (Topic) ซึ่งเรียกว่า Topic Node โดย ภายในประเด็นของเอกสาร จะประกอบไปด้วย คำสำคัญที่ความหมายใกล้เคียงกัน
- Topic Node จะเชื่อมต่อกันเป็นกราฟแบบสมบูรณ์ (Complete Graph)
- Topic Node จะถูกจัดอันดับโดยคำนวณจากค่าความเกี่ยวข้อง
- คำสำคัญจะถูกเลือก K ตัว จาก ประเด็นเอกสาร N ประเด็น

ซึ่งมีภาพรวมการทำงานดังแสดงในภาพที่ 2




**ภาพที่ 2** แสดงขั้นตอนการทำงานของการสกัดคำสำคัญโดยใช้ความหมาย

ขั้นตอนที่ 1 การเตรียมข้อมูล

เป็นขั้นตอนเตรียมข้อความให้อยู่ในรูปที่พร้อมใช้สำหรับขั้นตอนต่อไป ได้แก่

- การตัดคำ (Word Segmentation)
- การกำจัดคำหยุด (Stop Word Removing)
- การกำกับหน้าที่ของคำ (part-of-speech Tagging)

ในงานวิจัยนี้เรียกใช้ ไลบารี Spacy ในภาษาไพธอน เพื่อทำหน้าที่ตัดคำและกำกับหน้าที่ของคำ

ขั้นตอนที่ 2 การคัดเลือกคำคู่แข่ง

คำคู่แข่ง คือคำที่มีรูปแบบทางไวยากรณ์เหมือนคำสำคัญ ซึ่งมีโอกาสที่จะถูกเลือกไปเป็นคำสำคัญในลำดับต่อไป ซึ่งงานวิจัยได้กำหนดให้ คำคู่แข่ง คือคำ หรือ กลุ่มคำ ตั้งแต่ 2-5 คำ ที่ทำหน้าที่เป็น คำนาม คำกริยา คำคุณศัพท์ และ คำวิเศษณ์ ซึ่งเราสามารถรู้หน้าที่ของคำได้จาก ขั้นตอนที่ 1 การเตรียมข้อมูล ในกรณีที่ เป็นกลุ่มคำ งานวิจัยนี้ จะใช้กฎของการรวมกลุ่มคำ (Multiword Formation Rules) ที่ รวบรวมมาจากผู้เชี่ยวชาญทางด้านภาษาศาสตร์ ดังแสดงในตาราง 2

**ตารางที่ 2** กฎของการรวมกลุ่มคำ (Multiword Formation Rules)

| Pattern | Example |
|---|---|
| n + n | แม่ (mother) + น้ำ (water)<br>แม่น้ำ (river) |
| n + v | ห้อง (room) + นอน (sleep)<br>ห้องนอน (bedroom) |
| v + n | พัด (fan) + ลม (wind)<br>พัดลม (a fan) |
| v + v | กัน (prevent) + ชน (crash)<br>กันชน (bumper) |
| n + adj | น้ำ (water) + แข็ง (solid)<br>น้ำแข็ง (ice) |
| adj + n | หวาน (sweet) + ใจ (heart)<br>หวานใจ (sweetheart) |
| n + n + v | สาย (line) + ไฟ (fire) + พ่วง (tow)<br>สายไฟพ่วง (power cable) |
| n + v + n | คน (human) + ขับ (drive) + รถ (car)<br>คนขับรถ (driver) |
| n + n + n | ข้าว (rice) + ขา (leg) + หมู (pig)<br>ข้าวขาหมู (stewed pork leg on rice) |
| n + v + v | ใบ (leaf) + ขับ (drive) + ขี่ (ride)<br>ใบขับขี่ (driving license) |

ขั้นตอนที่ 3 การจัดกลุ่มคำคู่แข่ง

เนื่องจากในโครงการวิจัยนี้ เสนอการสกัดคำสำคัญที่ครอบคลุมประเด็นหลักของเอกสาร ดังนั้น เราจำเป็นต้อง กำหนดประเด็นเอกสาร ขึ้นมาก่อน โดยการรวบรวมกลุ่มคำคู่แข่งที่มีความหมายใกล้เคียงกันเข้ามาอยู่ในประเด็นเดียวกัน ซึ่งเราได้ตั้ง สมมุติฐานไว้ว่า “ประเด็นของเอกสารสามารถกำหนดได้จากกลุ่มของคำคู่แข่งที่มีความหมายใกล้เคียงกัน” เทคนิคที่ใช้ในการจัดกลุ่มคู่แข่ง สำหรับโครงการวิจัยนี้ จะใช้วิธีการของ อัลกอริทึม K-mean สำหรับจัดกลุ่มข้อมูล โดยนำมาประยุกต์ใช้ในการ จัดกลุ่มคำคู่แข่ง ดังนี้

พิจารณา กราฟความหมาย ซึ่งถ้า ในเอกสารมีจำนวนประเด็นของเอกสารมากกว่าหนึ่ง เราจะได้กลุ่มของกราฟตามจำนวนของประเด็นเอกสาร

กำหนดให้ k หมายถึง จำนวนประเด็นของเอกสาร

CTi หมายถึง ค่ากลางของกลุ่มของกราฟที่ i ซึ่งเลือกมาจากคำที่อยู่ใน Concept node ที่มี จำนวน link เชื่อมโยง (ทั้ง In-link และ out-link) มากที่สุด

คำนวณ ค่าความเกี่ยวข้องทางความหมาย (Semantic Relevant Score) ระหว่าง คำคู่แข่งแต่ละตัวกับค่ากลางของกลุ่มกราฟ k กลุ่ม โดยสมการที่คำนวณมีดังนี้

$$i = \underset{i=1..k}{\operatorname{argmax}} \ (SR(C, CT_i)) \qquad (1)$$

โดย i คือ กลุ่มของกราฟ (ประเด็นเอกสาร) ที่คำคู่แข่งมีค่าความเกี่ยวข้องทางความหมายมากที่สุด

$$SR(C, CT_i) = S * \frac{R}{L} \qquad (2)$$

โดย S คือ ค่าที่บอกว่า C และ $CT_i$ อยู่ในกลุ่มกราฟเดียวกันหรือไม่

$$S = \begin{cases} 1, & \text{เมื่อ } \mathbf{C} \text{ และ } \mathbf{CT_i} \text{ อยู่ในกลุ่มกราฟเดียวกัน} \\ 1/k, & \text{เมื่อ } \mathbf{C} \text{ และ } \mathbf{CT_i} \text{ อยู่ในกลุ่มกราฟต่างกัน} \end{cases} \qquad (3)$$

โดย R คือ ค่าน้ำหนักสะสมของ Conceptual Relation Node จาก C และ $CT_i$ โดยงานวิจัยนี้กำหนดค่าน้ำหนักของ Conceptual Relation Node แตกต่างกันตามความสำคัญ ดังตาราง 3

L คือ จำนวนของ Edge จาก C และ $CT_i$

**ตารางที่ 3** แสดงค่าน้ำหนักที่แตกต่างกันของ Conceptual Relation Node

| Conceptual Relation Node | ค่าน้ำหนัก |
|---|---|
| Agnt, obj | $W_1$ (1) |
| Loc, dur, pitm | $W_2$ (0.75) |
| Attr, manner | $W_3$ (0.5) |
| อื่น ๆ | $W_4$ (0.25) |

หมายเหตุ: ค่าน้ำหนักในตารางสามารถปรับเปลี่ยนได้ตามความเห็นชอบของผู้เชี่ยวชาญ ตัวเลขดังกล่าวเป็นค่าที่ใช้ในงานวิจัยนี้

ขั้นตอนที่ 4 การจัดอันดับประเด็นเอกสาร

เมื่อ คำคู่แข่ง ถูกจัดกลุ่มเรียบร้อยแล้ว ลำดับต่อไป เราจะทำการจัดอันดับความสำคัญของประเด็นเอกสาร ในส่วนนี้ เราประยุกต์ใช้เทคนิคของ TopicRank [17] นั่นคือ ประเด็นเอกสาร ที่มีคำคู่แข่งที่อยู่ตำแหน่งต้น ๆ ของเอกสารจะมีความสำคัญมากกว่าประเด็นเอกสาร ที่มีคำคู่แข่งที่อยู่ตำแหน่งท้าย ๆ ของเอกสาร สมการที่ใช้ในการจัดอันดับเป็นดังนี้

$$w_{i,j} = \sum_{c_i \in t_i} \sum_{c_j \in t_j} \text{dist}(c_i, c_j) \tag{4}$$

$$\text{dist}(c_i, c_j) = \sum_{b_i \in \text{boz}(c_i)} \sum_{b_j \in \text{boz}(c_j)} \frac{|b_i - b_j|}{1} \tag{5}$$

โดย $w_{i,j}$ คือ ค่าน้ำหนักของ Edge ที่เชื่อมระหว่าง ประเด็นเอกสาร (Topic)

Pos($c_i$) คือ ตำแหน่งของ คำคู่แข่ง $C_i$ โดยนับเริ่มจากต้นเอกสาร

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{6}$$

ขั้นตอนนี้ เราจะได้ผลลัพธ์ คือ ประเด็นของเอกสาร ที่มีอันดับสูงสุด k ประเด็น

ขั้นตอนที่ 5 การสกัดคำสำคัญ

เมื่อเราได้ประเด็นหลักของเอกสารแล้วจากขั้นตอนที่ 4 ลำดับต่อไปคือเลือก คำสำคัญ ออกมาจากแต่ละประเด็น ซึ่งในที่นี้ เราจะเลือก คำคู่แข่ง ที่ปรากฏในเอกสาร ในตำแหน่งก่อนที่สุดเป็น คำสำคัญ (Keyword)

**ผลการศึกษา**

การสกัดคำสำคัญโดยใช้ความหมายของเอกสาร จะเปรียบเทียบผลการทดลองกับ งานวิจัยของ

- KEA [3] เป็นตัวแทนของงานวิจัยที่สกัดคำสำคัญโดยใช้หลักทางสถิติ คำนวณจากคลังเอกสาร ที่ได้รวบรวมไว้ ซึ่งใช้การพิจารณาเชิงไวยากรณ์ในการเลือกคำสำคัญ และ
- TextRank [17] เป็นตัวแทนของงานวิจัยที่สกัดคำสำคัญที่ใช้หลักของ Graph-based Keyword Extraction โดยแต่ละ Node คือคำคู่แข่ง ที่อยู่ในเอกสาร ที่ถูกเลือกมาโดยพิจารณาคุณลักษณะทางไวยากรณ์ และจากนั้นจัดอันดับคำคู่แข่ง โดยใช้ความเชื่อมโยงของ Node ในการคำนวณค่าคะแนนความเกี่ยวข้อง คำคู่แข่งที่มีคะแนนสูงสุด N จำนวนถูกเลือกออกมาเป็นคำสำคัญ
- TopicRank [18] เป็นตัวแทนของงานวิจัยที่สกัดคำสำคัญที่ใช้หลักของ Graph-based Keyword Extraction โดยแต่ละ Node คือ ประเด็นของเอกสารที่ได้จากการคัดเลือกคำคู่แข่งตามคุณลักษณะทางไวยากรณ์ แล้วจัดกลุ่มของคำคู่แข่งแล้วกำหนดเป็นประเด็น จากนั้นจัดอันดับประเด็นเอกสาร เรียงตามความสำคัญของคำที่อยู่ในประเด็น จากนั้นเลือกคำสำคัญ จากแต่ละประเด็นออกมา

**1. การวัดผลงานวิจัย**

ในการประเมินประสิทธิภาพของงานวิจัย เราจะใช้ ค่าความแม่นยำ ค่าความระลึกได้ และ ค่า F1-Measure ซึ่งคำนวนณจากสมการดังต่อไปนี้

$$Pr\,e\,cision(P) = \frac{TP}{TP+TN} \quad (7)$$

$$Re\,c\,all(R) = \frac{TP}{TP+FN} \quad (8)$$

$$F1\_measure = \frac{2\,PR}{P+R} \quad (9)$$

โดย ค่าของ TP, TN, FP, FN พิจารณาดังตารางที่ 4

**ตารางที่ 4** แสดง Confusion Matrix ระหว่างมนุษย์และคอมพิวเตอร์

| | System say "Yes" | System say "No" |
|---|---|---|
| Human say "Yes" | True Positive (TP) | False Negative (FN) |
| Human say "No" | False Positive (FP) | True Negative (TN) |

บทความวิจัยนี้ ประเมินประสิทธิภาพของ วิธีที่เสนอ ด้วยการใช้ข้อมูลชุดเดียวกัน จำนวน 524 เอกสาร โดยทดลองกับ วิธีการสกัดคำสำคัญจาก งานวิจัยที่ผ่านมา 3 ระบบ ได้แก่ KEA [3] ซึ่งเป็นสกัดคำสำคัญที่พิจารณาความคล้ายกันของจากคุณลักษณะทางไวยากรณ์ของคำซึ่งจะพิจารณาแยกแต่ละคำ TextRank [17] เป็นการสกัดคำสำคัญที่ใช้หลักของ Graph-based Keyword Extraction แสดงให้เห็นความเชื่อมโยงของคำ และ TopicRank [18] ใช้หลักของ Graph-based Keyword Extraction เช่นกันแต่มีการจัดกลุ่มของคำตามประเด็นของเอกสาร โดยพิจารณาจากการเกิดคู่กันของคำ ซึ่งประสิทธิภาพของวิธีที่เสนอที่ใช้กราฟมโนทัศน์แสดงความสัมพันธ์ทางความหมายของทั้งประโยค ดังแสดงในตาราง 5 มีประสิทธิภาพโดยรวมสูงกว่าอย่างมีนัยสำคัญ ส่วนตารางที่ 6 แสดงผลของการสกัดคำสำคัญ ด้วยวิธีที่เสนอ เมื่อกำหนดจำนวนของคำสำคัญต่างกัน โดยกำหนดที่ 5 และ 10 ตามลำดับ

**วิจารณ์**

จากผลของการทดลองที่ใช้ความหมายในการสกัดคำสำคัญ สามารถเพิ่มประสิทธิภาพของการสกัดคำสำคัญของเอกสารได้อย่างมีนัยสำคัญ แต่อย่างไรก็ตาม ค่าความแม่นยำ ค่าความระลึก และ ค่า F1-measure ที่ได้ยังไม่สูงนักถ้าจะเอาไปใช้งานจริง สาเหตุที่เป็นเช่นนั้น เนื่องจาก ข้อจำกัด ดังนี้

**ตารางที่ 5** ประสิทธิภาพของงานวิจัยเทียบกับงานวิจัยที่เกี่ยวข้อง

| Keyphrase Approaches | Performance | | |
|---|---|---|---|
| | % of Recall | % of Precision | F1-measure |
| KEA | 32 | 32 | 32 |
| TextRank | 37 | 37 | 37 |
| TopicRank | 38 | 38 | 38 |
| Proposed Approach* | 53 | 40 | 40 |

**ตารางที่ 6** ประสิทธิภาพของงานวิจัยเมื่อกำหนดจำนวนคำสำคัญแตกต่างกัน

| Number of desired keyphrases | Performance | | |
|---|---|---|---|
| | % of Recall | % of Precision | F1-measure |
| 5 | 38 | 47 | 32 |
| 10 | 53 | 33 | 37 |

(1) ในการใช้งานจริง ผู้ใช้จะกำหนดคำสำคัญขึ้นจากความเข้าใจความหมายของเนื้อหา ซึ่งในหลายครั้งจะกำหนดคำขึ้นมาเอง ซึ่งอาจจะไม่มีคำเหล่านั้นปรากฏอยู่ในเอกสาร เราเรียกการ กำหนดคำสำคัญแบบนี้ว่า Keyword Assignment ในขณะที่คอมพิวเตอร์ จะพิจารณาคำสำคัญ โดยดูจากคำ หรือ กลุ่มคำ ที่ปรากฏอยู่ในเอกสาร และจะพิจารณาคำที่อยู่ติดกันเป็นหลัก เมื่อนำมาประเมินประสิทธิภาพด้วย ค่าความแม่นยำ ค่าความระลึก และ ค่า F1-Measure ที่จะเปรียบเทียบโดยใช้ การเปรียบเทียบชุดอักษร ทำให้ค่า True Positive น้อยกว่าที่ควรจะเป็น ยกตัวอย่างเช่น

คำสำคัญ (ผลเฉลย) “Ministry of Finance”
คำสำคัญที่สกัดได้ “Finance Ministry”

(2) ในส่วนของการสกัดคำสำคัญจากเอกสารภาษาไทยนั้น มีข้อจำกัดดังนี้

- ส่วนของการสร้างกลุ่มคำ เพื่อนำไปใช้ในการสร้าง คำคู่แข่ง ในงานวิจัยนี้ ใช้กฎ ที่สร้างจากคำแนะนำของผู้เชี่ยวชาญในการสกัดรูปแบบ โดยจะแบ่งกลุ่มคำออกเป็น 3 ประเภท คือ

1) กลุ่มคำนาม (Noun phrase)
2) กลุ่มคำนามเฉพาะ (Named Entity)
3) กลุ่มคำกริยาเรียง (Serial Verb)

และจำนวนของกลุ่มคำที่ใช้ในงานวิจัยนี้ เป็นไปได้ตั้งแต่ 2-5 คำ ดังนั้นในกรณีที่จำนวนคำมากกว่าที่กำหนด เช่น คำที่เป็นชื่อเฉพาะ เช่น ชื่อองค์กร หรือ สถานที่ อาจทำให้ผิดพลาดได้

- ส่วนของการเปรียบเทียบกับกราฟความหมาย เนื่องจากเครื่องมือสร้างกราฟความหมาย ของงานวิจัยนี้ จำเป็นต้องใช้ ตัวแจกแจงไวยากรณ์แบบพึ่งพา ซึ่งในปัจจุบันยังไม่รองรับภาษาไทย โดยงานวิจัยนี้ แก้ปัญหาด้วยการ แปลเอกสารเป็นภาษาอังกฤษ โดยใช้ไลบรารี่ Spacy และปรับให้ถูกต้องโดยตามความหมายโดยผู้เชี่ยวชาญ จากนั้นนำไปสร้าง

เป็นกราฟความหมาย ดังนั้น ความผิดพลาดของการสกัดคำสำคัญ ส่วนหนึ่งมาจากความผิดพลาดจากการแปลภาษา

## สรุป

บทความนี้นำเสนอหลักการทางการประมวลผลภาษาธรรมชาติ สำหรับสร้างกราฟความหมาย ซึ่งเป็นองค์ความรู้ในการสกัดคำสำคัญจากเอกสารโดยอัตโนมัติ ซึ่งเป็นส่วนสำคัญของเครื่องมือค้นหาข้อมูลเชิงลึก โดยผลการทำลองมีประสิทธิภาพสูงกว่างานวิจัยที่ผ่านมาอย่างมีนัยสำคัญ โดยผลที่ได้จากการสกัดคำสำคัญแบบอัตโนมัติ มีค่าความแม่นยำ ร้อยละ 40 ค่าความครบถ้วน ร้อยละ 53 และ F1-Measure ที่ ร้อยละ 40 ตามลำดับ สำหรับแนวทางในการวิจัยต่อยอดการนำเอากราฟมโนทัศน์มาใช้ในการสอนให้คอมพิวเตอร์เข้าใจภาษามนุษย์ได้นั้น อาจจะต้องขยายบริบทของการเปรียบเทียบให้กว้างขึ้นกว่าระดับประโยค เช่น เปรียบเทียบในระดับของย่อหน้า หรือ ระดับเอกสาร เป็นต้น นอกจากนี้เครื่องมือค้นหาเชิงลึกมีบทบาทสำคัญในการอำนวยความสะดวกในการค้นหาข้อมูล การตัดสินใจ และการวิจัยเฉพาะโดเมนในอุตสาหกรรมและโดเมนต่างๆ โดยนำเสนอความสามารถในการค้นหาเฉพาะทางที่ปรับให้เหมาะกับความต้องการและความชอบเฉพาะของผู้ใช้ เช่น งานการค้นหาด้านกฎหมาย การค้นหาข้อมูลทางการแพทย์ เป็นต้น ซึ่งจำเป็นต้องมีความเข้าใจความหมายและหาคำตอบได้ถูกต้อง

## กิตติกรรมประกาศ

บทความนี้เป็นส่วนหนึ่งของโครงการวิจัยหัวข้อ “โครงการพัฒนาเครื่องมือค้นหาบนอินเตอร์เน็ตแบบเชิงลึก” ภายใต้ชุดโครงการพัฒนาโครงสร้างพื้นฐาน สำหรับเครื่องมือค้นหาบนอินเตอร์เน็ต และระบบสืบค้นข้อมูลเชิงลึก เลขที่ 66807 งบประมาณบูรณาการวิจัยและนวัตกรรม ปีงบประมาณ พ.ศ. 2562 โดยได้รับการสนับสนุนจากสำนักงานการวิจัยแห่งชาติ (วช.)

## เอกสารอ้างอิง (References)

1. Page, Larry. PageRank: Bringing Order to the Web. Stanford Digital Library Project [Internet]. 2002 [cited 2020 August 31]. Available from: http://infolab.stanford.edu/~page/papers/pagerank/index.htm
2. Microsoft Inc. Microsoft's New Search at Bing.com Helps People Make Better Decisions. [Internet]. 2009. [cited 2020 August 31]. Available from:

https://news.microsoft.com/2009/05/28/microsofts-new-search-at-bing-com-helps-people-make-better-decisions/

3. Wu YB, Li Q, Bot RS, Chen X. Domain-Specific Keyphrase Extraction. Proceedings of the 14th ACM international conference on Information and knowledge management; Bremen, Germany, 2005;283-4.
4. Turney P. Learning algorithms for keyphrase extraction. Information Retrieval 2000;2:303–36.
5. Tomokiyo T, Hurst M. A Language Model Approach to Keyphrase Extraction. Proceedings of the ACL Workshop on Multiword Expressions; Sapporo, Japan, 2003;33–40.
6. Barker K, Cornacchia N. Using Noun Phrase Heads to Extract Document Keyphrases. Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence; Montreal, Quebec, Canada, 2000:40–52.
7. Liu F, Pennell D, Liu F, Liu Y. Unsupervised approaches for automatic keyword extraction using meeting transcripts. Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics; Boulder, Colorado, 2009:620–8.
8. Matsuo Y, Ishizuka M. Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 2004:157-69.
9. Grineva M, Grinev M, Lizorkin D. Extracting key terms from noisy and multitheme documents. Proceedings of the 18th International Conference on World Wide Web; New York, NY, 2009: 661–70
10. Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of the 10th Research on Computational Linguistics International Conference; Taipei, Taiwan, 1997:19-33.
11. Lin D. An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning; San Francisco, CA, 1998:296-304.

12. Wu Z, PalmerM. Verb semantics and lexical selection. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics; Las Cruces New Mexico, 1994:133-8.
13. BrightEdge.com. What is vertical search optimization or VSO?. [Internet]. 2018. [cited 2020 August 31]. Available from: https://www.brightedge.com /blog/vso-vertical-search-optimization)
14. Long B, Chang Y. Relevance Ranking for Vertical Search Engines; Morgan Kaufmann, 2014.
15. Kim SN, Medelyan O, Kan MY, Baldwin T. SemEval-2010 Task 5 : Automatic Keyphrase Extraction from Scientific Articles. Proceedings of the 5th International Workshop on Semantic Evaluation; Uppsala, Sweden; 2010: 21-6.
16. Sowa JF. Conceptual Graph Standard and Extension. Conceptual Structures: Theory, Tools and Applications. Proceedings of the 6th International Conference on Conceptual Structures; Montpellier, France. August 10-12, 1998:3-14.
17. Mihalcea R, Tarau P. Textrank: Bringing order into text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing; Barcelona, Spain, 2004: 404–11.
18. Bougouin A , Boudin F , Daille B. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP); Nagoya, Japan, 2013:543-51.