

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของงานวิจัย

การค้นคืนสารสนเทศ (Information Retrieval) เป็นสิ่งที่จำเป็นอย่างมากสำหรับการดำเนินชีวิตของมนุษย์ เนื่องจากข้อมูลต่างๆ มีอยู่เป็นจำนวนมาก ถ้าต้องให้เรา来选择สารสนเทศที่ต้องการด้วยแรงงานคน ก็จะทำให้เสียเวลามาก และบางครั้งอาจไม่ได้ข้อมูลตามที่ต้องการด้วย ซึ่งต่อมาได้พยายามนำเอาคอมพิวเตอร์เข้ามาช่วยในการจัดเก็บข้อมูล โดยจะรับคำถาม (query) จากผู้ใช้ แล้วคอมพิวเตอร์จะพยายามค้นหาข้อมูลในฐานข้อมูลที่ตรงกับคำถามที่ป้อนให้ ในระยะแรกๆ การค้นคืนสารสนเทศจากฐานข้อมูล ก็จะคล้ายกับ การใช้ภาษา SQL (Structured Query Language) ในการเรียกค้นข้อมูลจากฐานข้อมูล

ในการเปรียบเทียบระหว่างคำถามและข้อมูลซึ่งอยู่ในรูปของเอกสารโดยระบบการค้นคืนสารสนเทศนั้นจะใช้วิธีการวัดความคล้าย (Similarity) ของเอกสารโดยดูจากความคล้ายกันกับคำที่ผู้ใช้ให้เป็นตัวอย่าง (Query Example) ซึ่งในระยะแรกจะใช้การเปรียบเทียบรูปแบบการเขียนว่าเขียนเหมือนกันหรือไม่ เช่น คำถามต้องการหาคำว่า “Operate” ระบบค้นคืนสารสนเทศจะค้นหาเอกสารที่มีคำว่า “Operate” ปรากฏอยู่แล้วแสดงผลให้กับผู้ใช้ ด้วยวิธีนี้มีข้อจำกัดตรงที่คำโดยทั่วไป โดยเฉพาะคำในภาษาอังกฤษสามารถเปลี่ยนรูปได้ เช่น “Operating” “Operation” และ “Operated” เป็นต้น เอกสารต่างๆ ที่มีคำเหล่านี้อยู่ อาจจะไม่ถูกเลือกเพราะถ้าเปรียบเทียบกันแบบตัวอักษรต่อตัวอักษร แล้ว คำถามกับคำเหล่านี้ต่างกัน ต่อมาได้พยายามเพิ่มความฉลาดให้กับระบบการค้นคืนสารสนเทศ ด้วยการนำเอาวิธีการทางด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) เข้ามาช่วยจัดการเพื่อตีความหมายของคำถามที่ผู้ใช้ป้อนเข้ามา ซึ่งสามารถตอบได้ว่า “Operating” “Operation” และ “Operated” เป็นคำเดียวกันกับ “Operate” ดังนั้น เราสามารถสืบค้นสารสนเทศได้ครอบคลุมมากขึ้น

แต่อย่างไรก็ตามการอธิบายในสิ่งที่มีความหมายเดียวกัน เราสามารถใช้คำมากกว่า 1 คำ จึงมีผลทำให้ผลการค้นคืนเอกสารและข้อมูลที่ได้ไม่ครอบคลุมความต้องการของผู้ใช้ นั่นคือระบบการค้นคืนสารสนเทศซึ่งพิจารณาแค่ความเหมือนของคำที่พิจารณาเฉพาะรูปแบบที่เขียนนั้น จะไม่สามารถให้ผลครอบคลุมความต้องการของผู้ใช้งานได้ ดังนั้นเพื่อลดปัญหาดังกล่าวและให้

การค้นคืนสารสนเทศได้ผลครอบคลุมและตรงหรือใกล้เคียงกับความต้องการของผู้ใช้มากที่สุด จึงมีการนำเอาความหมายของคำมาพิจารณา ซึ่งเรียกว่า ระบบการค้นคืนข้อมูลสารสนเทศที่พิจารณาเปรียบเทียบจากความคล้ายกันในเชิงความหมาย (Semantic-Based Similarity)

ถึงแม้ว่าการค้นคืนสารสนเทศโดยวิธีพิจารณาเปรียบเทียบความสัมพันธ์กันในเชิงความหมายนั้นจะค้นคืนเอกสารได้ครอบคลุมตรงตามต้องการของผู้ใช้มากกว่า โดยจะพิจารณาคำที่อธิบายในสิ่งที่มีความหมายเดียวกัน เราสามารถใช้คำมากกว่า 1 คำ อธิบายได้ เช่น

- (1) “car” และ “motorcar” ซึ่งมีความหมายเดียวกันแต่มีต้นกำเนิดมาจากคนละที่ (Synonym)
- (2) “car” และ “motor vehicle” มีความหมายคล้ายกันโดยที่ “motor vehicle” มีความหมายทั่วไปกว่า (Hypernym)
- (3) “car” และ “cruiser” มีความหมายคล้ายกันโดยที่ “cruiser” มีความหมายเฉพาะเจาะจงกว่า (Hyponym)

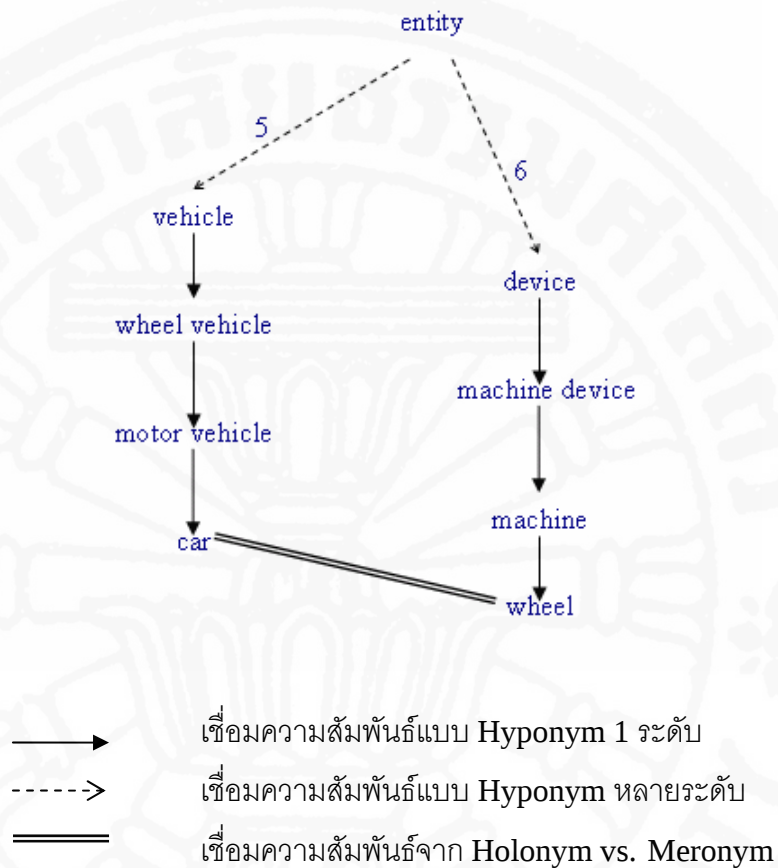
แต่อย่างไรก็ตามผลที่ได้จากการค้นคืนสารสนเทศโดยวิธีพิจารณาเปรียบเทียบความสัมพันธ์กันในเชิงความหมายนั้นจะได้ผลจากการค้นคืนเป็นจำนวนมาก ทำให้ผู้ใช้ต้องเสียเวลาในการคัดเลือก รวมทั้งในการพิจารณาความสัมพันธ์เชิงความหมายทั้ง 3 รูปแบบในข้างต้นนั้นไม่ได้มีการให้ความสำคัญของความสัมพันธ์ในแต่ละรูปที่แตกต่างกัน ทำให้เกิดปัญหาในการจัดลำดับการแสดงผล

นอกจากนี้ ความสัมพันธ์ในเชิงความหมายที่นอกเหนือจากความสัมพันธ์เชิงความหมายทั้ง 3 รูปแบบในข้างต้นไม่ได้ถูกนำมาพิจารณาเปรียบเทียบด้วย ทำให้ผลการค้นที่ได้ไม่ครอบคลุมความสัมพันธ์ในเชิงความหมายรูปแบบเหล่านี้ เช่น

- (1) “car” และ “train” มีความหมายคล้ายกันโดยที่ “car” มีความหมายเป็นส่วนประกอบของ “trains” (Holonym)
- (2) “car” และ “wheel” มีความหมายคล้ายกันโดยที่ “car” มี “wheel” เป็นส่วนประกอบ (Meronym)

ภาพที่ 1.1

แสดงความสัมพันธ์โดยพิจารณาความสัมพันธ์ที่เป็น “is-a” ในโครงข่ายเวิร์ดเน็ต



จากงานวิจัยที่ผ่านมาในจะพิจารณาความสัมพันธ์ที่เป็น “is-a” หรือ ความสัมพันธ์แบบ Synonym Hypernym และ Hyponym ซึ่งจากภาพที่ 1.1 จะเห็นได้ว่าการวัดความสัมพันธ์ของ “car” กับ “wheel” หากพิจารณาจากระยะห่างระหว่างคำทั้งสองโดยอาศัยความสัมพันธ์ที่เป็น “is-a” นั้น จะมีระยะห่างกันเท่ากับ 17 นั่นคือมีความสัมพันธ์กันน้อยมาก ทั้งที่ “wheel” มีความสัมพันธ์เป็น “part-of” คือเป็นส่วนประกอบของ “car”

ดังนั้น วิทยานิพนธ์นี้ จึงเสนอการให้น้ำหนักความสำคัญที่แตกต่างกันในการหาความสัมพันธ์กันของคำในเชิงความหมาย และการเสนอการพิจารณาความสัมพันธ์เชิงความหมายในรูปแบบที่เป็น Holonym และ Meronym ซึ่งเป็นความสัมพันธ์ “part-of”, ความสัมพันธ์ “member-is-a-kind-of”, ความสัมพันธ์ “member-is-a-part-of” นอกเหนือจากความสัมพันธ์ที่เป็น “is-a” ซึ่งประกอบด้วย Synonym Hypernym และ Hyponym โดย

พิจารณาจากความหมายในระดับประโยคซึ่งในการพิจารณาในระดับประโยคจะพิจารณาความหมายของทั้งประโยคที่เป็นคำแปลของคำที่พิจารณา ซึ่งการพิจารณาความหมายของประโยคนั้น จำเป็นจะต้องแทนความหมายของประโยคให้อยู่ในรูปที่คอมพิวเตอร์เข้าใจเสียก่อน ซึ่งในวิทยานิพนธ์นี้เลือกใช้การแทนความหมายด้วยกราฟเชิงความคิด (Conceptual Graph) จากนั้น คำนวณหาความสัมพันธ์จากผลที่ได้จากการเปรียบเทียบกันของกราฟเชิงความคิดซึ่งใช้แทนความหมายของประโยค

1.2 วัตถุประสงค์ของงานวิจัย

1. เพื่อเสนอแนวทางในการเปรียบเทียบหาความสัมพันธ์ทางความหมายระหว่างคำ
2. เพื่อเสนอวิธีการในการแทนความหมายของประโยค

1.3 ขอบเขตของงานวิจัย

งานวิจัยนี้เสนอวิธีการในการเปรียบเทียบความสัมพันธ์เชิงความหมายระหว่างคำ โดยมีขอบเขตของงานวิจัยดังต่อไปนี้

- 1.3.1 คำต่างๆที่ใช้เป็นภาษาอังกฤษ เพราะภาษาอังกฤษไม่มีปัญหาในเรื่องของการแบ่งคำและประโยค
- 1.3.2 ในการเปรียบเทียบความหมายของคำ จะใช้ข้อมูลจาก ฐานข้อมูลของคำที่ชื่อว่า “WordNet” เป็นหลัก เนื่องจากเป็นฐานข้อมูลที่ให้บริการโดยไม่คิดค่าใช้จ่าย สามารถดาวน์โหลดได้จากอินเทอร์เน็ต และเป็นฐานข้อมูลที่ใช้กันในงานวิจัยทั่วไป
- 1.3.3 ประโยคที่ใช้ในการเปรียบเทียบความสัมพันธ์เชิงความหมายใช้ประโยคซึ่งเป็นคำแปล (Glossary) ของคำที่ต้องการเปรียบเทียบ
- 1.3.4 ในการประเมินผลจะเปรียบเทียบผลการวัดความสัมพันธ์เชิงความหมายของคำ โดยใช้ความหมายของประโยคกับผู้ประเมิน

1.4 ผลที่ได้รับ

วิทยานิพนธ์นี้ได้เสนอผลของการวิจัยดังนี้

- 1.4.1 วิธีการเปรียบเทียบความสัมพันธ์ทางความหมายระหว่างคำ ที่ให้ผลออกมาเป็นตัวเลขสามารถเปรียบเทียบกันได้
- 1.4.2 แนวทางในการแทนความหมายของประโยค สามารถนำไปประยุกต์ใช้กับงานอื่นๆ ได้ เช่น การแปลเอกสาร (Machine Translation) การสรุปใจความสำคัญ (Text Summarization) และ การจำแนกเอกสาร (Text Categorization) เป็นต้น

1.5 รายละเอียดของวิทยานิพนธ์

วิทยานิพนธ์ฉบับนี้ประกอบด้วยเนื้อหา 4 ส่วนด้วยกันคือ บทนำซึ่งเป็นบทที่ 1 จะกล่าวถึงความเป็นมาของปัญหา วัตถุประสงค์ของงานวิจัย ขอบเขตของงานวิจัย และผลที่จะได้รับ ในบทที่ 2 จะกล่าวถึงทฤษฎีต่างๆ เกี่ยวข้องกับงานวิจัย รวมทั้งงานวิจัยอื่น ๆ ที่เกี่ยวข้อง ในบทที่ 3 เป็นการอธิบายวิธีการดำเนินงานวิจัย ขั้นตอนการดำเนินงาน และแผนการดำเนินงาน ส่วนผลการทดลองและประเมินผลจะอยู่ในบทที่ 4 บทสุดท้ายคือบทที่ 5 เป็นการสรุปผลการศึกษารวมทั้งรายละเอียดข้อเสนอแนะ