บทที่ 1

บทนำ

1. บทนำต้นเรื่อง

คำหรือวลีสำคัญของเอกสารที่ดีและเหมาะสมที่จะนำมาเป็นดรรชนีคำค้นของเอกสารนั้น เป็นเครื่องมือสำคัญในการสืบค้นสารสนเทศ (Information Retrieval) ทั้งนี้ ดรรชนีคำหรือวลีสำคัญ (Keyword & Keyphrase Index) เป็นลักษณะของดรรชนีที่ครอบคลุมคำสำคัญทั้งหลายที่ปรากฏในเอกสาร โดยเอกสาร (source document) ที่ใช้ทำดรรชนี อาจใช้ชื่อเรื่องเอกสาร (Title) บทคัดย่อหรือสาระสังเขป (abstract) หรือตัวเอกสาร (full text) อย่างใดอย่างหนึ่งหรือมากกว่าหนึ่งอย่าง การทำดรรชนีโดยระบบคอมพิวเตอร์เป็นเรื่องยากและซับซ้อน การหาคำหรือวลีสำคัญในตัวเอกสารหรือการทำดรรชนีจากตัวเอกสารนั้น สามารถให้ประสิทธิผลมากกว่าการใช้ชื่อเรื่อง หรือบทคัดย่อ แต่อาจทำให้เสียเวลาประมวลผลมากและผลประโยชน์ที่ได้รับอาจไม่คุ้มค่า สำหรับการใช้ชื่อเรื่องรวมกับบทคัดย่อก็สามารถให้ประสิทธิผลเท่ากับการใช้เอกสาร ขณะเดียวกันยังให้ประสิทธิภาพในการสืบค้นสารสนเทศตามลักษณะของภาษาธรรมชาติของเอกสาร (free text search) อีกด้วย

โดยงานวิจัยนี้จะเป็นการพิจารณาข้อมูลทั้งเอกสาร (full document) สำหรับการค้นหาและระบุคำหรือวลีสำคัญ ที่สามารถเป็นตัวแทนเอกสารที่ดีในการทำดรรชนีคำค้น การทำดรรชนีแบ่งได้ 2 ประเภท ประเภทแรก คือ การดึงคำหรือวลีจากเอกสารโดยตรง (Extraction Method) ประเภทที่สอง คือ การกำหนดคำหรือวลีจากเนื้อหาเอกสาร (Assignment Method) โดยในงานวิจัยนี้จะเป็นการอาศัยเทคนิคในแบบแรกในการค้นหาคำหรือวลีสำคัญของเอกสารที่ผ่านขั้นตอนจากกระบวนการตัดคำ และก่อนจะทำการกำหนดคำหรือวลีสำคัญนั้น จะมีการกลั่นกรองคำหยุด คือ คำที่ไม่เหมาะสมจะนำมาเป็นตัวแทนเอกสารก่อน จึงจะนำกลุ่มคำหรือวลีเหล่านี้มาเป็นตัวแทนเอกสารในการทำดรรชนีคำค้น

2. ความสำคัญและที่มาของงานวิจัย

การศึกษานี้ เป็นการออกแบบแนวคิดที่จะช่วยในการสกัดหรือดึงคำหรือวลีที่ดีและมีความเหมาะสมที่สามารถนำมาเป็นตัวแทนของเอกสารแต่ละฉบับสำหรับการสร้างดรรชนีคำค้นของระบบการค้นหา และด้วยในปัจจุบันนั้น ปริมาณของสารสนเทศที่มีอยู่อย่างมากมายใน

เครือข่ายอินเตอร์เน็ต ก่อให้เกิดการพัฒนาระบบการค้นหาสารสนเทศที่สามารถตอบสนองการบริการด้านค้นหา หรือค้นคืนสารสนเทศที่มีอยู่ในเครือข่ายอย่างชาญฉลาด ด้วยเนื่องจากระบบการค้นคืนสารสนเทศที่ดีเป็นสิ่งจำเป็นอย่างมากสำหรับการให้บริการสารสนเทศบนเครือข่ายอินเตอร์เน็ต

ระบบการค้นคืนเอกสารบางระบบนั้นไม่อาจค้นหาเอกสารได้ตรงตามความต้องการของผู้ใช้ อาจเป็นเพราะเอกสารขาดการทำรายการที่ดีหรือการแทนเอกสารด้วยคำหรือวลีสำคัญที่ไม่สามารถเป็นตัวแทนเอกสารได้ดีพอ โดยที่ระบบดังกล่าวไม่สามารถประเมินได้ว่าเอกสารที่ค้นคืนออกมานั้นเป็นเอกสารที่ตรงกับความต้องการของผู้ใช้หรือไม่ บ่อยครั้งที่เอกสารที่ค้นคืนมานั้นไม่ได้เป็นเอกสารที่ต้องการ โดยอาจเป็นเพราะระบบค้นหา (Search Engine) ในปัจจุบันจะทำการค้นหาคำทุกคำที่ตรงกับคำค้น (Keyword & Keyphrase Search) ที่ผู้ใช้ป้อนออกมาทั้งหมดซึ่งอาจมากเกินความต้องการได้

การควบคุมความถูกต้องของภาษาก็เป็นสิ่งสำคัญในการค้นคืน เพราะภาษาบางภาษามีเอกลักษณ์หรือลักษณะเฉพาะ โดยเฉพาะภาษาไทยนั้นอุปสรรคของการค้นคืน ก็คือการแยกคำจากชุดสายอักขระ (Segmentation) ทั้งยังมีปัญหาเรื่องของคำหยุดในภาษาไทย โดยจำนวนคำหยุดของไทยมีผลกระทบกับประสิทธิภาพของของการประมวลผลในด้านของค่าความถูกต้อง รวมถึงในด้านการใช้เนื้อที่ในการเก็บหน่วยความจำสำรอง อีกทั้งยังมีปัญหาในเรื่องการกำหนดหน่วยคำซึ่งคำบางคำอาจเป็นคำศัพท์ที่เกิดขึ้นจากหน่วยคำมากกว่าหนึ่งหน่วย หรือที่เรียกว่าอาจใช้วลี โดยวลีเกิดจากนำคำมากกว่าหนึ่งคำมารวมกันสำหรับการแทนคำที่มีสิ่งใดสิ่งหนึ่งที่อาจเป็นรูปธรรมหรือนามธรรม ก่อให้เกิดความกำกวมซึ่งเป็นปัญหาหลักอย่างหนึ่งของการตัดคำไทย ด้วยเหตุนี้ การหาขอบเขตหรือการตัดคำเป็นสิ่งที่จำเป็นที่จะต้องทำเป็นอันดับแรกสำหรับการสกัดคำหรือวลีสำคัญหรือในงานด้านประมวลผลภาษาธรรมชาติ

การสกัดคำหรือวลีสำคัญที่เป็นตัวแทนของเอกสารนั้น เป็นการพิจารณาค่าความสำคัญของคำหรือวลี โดยที่หลักการเบื้องต้นในการให้ความสำคัญของคำ คือ หลักการของระบบการให้น้ำหนักของคำ (Term Weighting) และรวมถึงเป็นกระบวนการวิเคราะห์เนื้อหาของเอกสาร ในอดีตนั้นการพิจารณาเลือกวลีสำคัญส่วนใหญ่จะกระทำโดยผู้เชี่ยวชาญ และการที่ผู้เชี่ยวชาญพิจารณาเลือกคำหรือวลีสำคัญโดยพิจารณาเอกสารทั้งหมดที่ละฉบับเป็นงานที่ทำได้ยากและต้องใช้เวลามาก

ด้วยความจำเป็นดังที่ได้กล่าวมานี้ ผู้วิจัยจึงได้เล็งเห็นถึงความจำเป็นในการตอบสนองต่อประสิทธิภาพในการจัดการด้านการค้นคืนสารสนเทศ โดยงานวิจัยนี้จึงมีวัตถุประสงค์ที่จะกล่าวถึงพื้นฐานแนวคิดของการตัดแยกคำหรือวลีไทย การให้ค่าความสำคัญหรือ

ระบบการให้น้ำหนักของคำหรือวลี เพื่อการสกัดคำหรือวลีที่สำคัญเป็นตัวแทนของเอกสารสำหรับระบบการค้นคืนที่ชาญฉลาดและมีประสิทธิภาพ

3. การตรวจเอกสาร (งานวิจัยที่เกี่ยวข้อง)

3.1 จากงานของ ถิรนันท์ ดำรงค์สอน และพีรวัฒน์ วัฒนพงศ์ (2545) มีการนำเสนอการสกัดวลีสำคัญแบบอัตโนมัติโดยใช้โครงข่ายประสาทเทียม โดยงานวิจัยนี้นำเสนอกระบวนการสกัดวลีสำคัญจากเอกสารภาษาอังกฤษแบบอัตโนมัติโดยใช้โครงข่ายประสาทเทียม โดยกระบวนการทำงานเริ่มจากสร้างแบบจำลองโครงข่ายประสาทเทียมสำหรับพิจารณาเลือกวลีสำคัญ โดยเรียนรู้จากเอกสารตัวอย่างที่มีวลีสำคัญของผู้เขียนกำกับอยู่ จากนั้นจะนำแบบจำลองที่ได้ไปใช้พิจารณาเลือกวลีสำคัญจากเอกสารใหม่ ทุกๆ วลีในเอกสารจะถูกเรียงลำดับความสำคัญโดยพิจารณาจากความถี่และตำแหน่งของวลี วลีที่ถูกจัดอยู่ในลำดับต้นๆ จะถูกเลือกเป็นวลีสำคัญของเอกสารนั้น โดยผลการทดสอบของงานวิจัยนี้พบว่า แบบจำลองโครงข่ายประสาทเทียมสามารถเลือกวลีสำคัญของเอกสารได้ถูกต้องโดยเฉลี่ยประมาณ 40%

3.2 จากงานของ รัตติกร วรากูลศิริพันธ์ และคณะ (1989) มีการนำเสนอเทคนิคในการเลือกประโยคที่ถูกต้องหลังจากการแยกแยะหน่วยคำ โดยอาศัยความถี่ของการใช้คำต่าง ๆ ในชีวิตประจำวันที่ถูกต้องตามหลักไวยากรณ์ของภาษาไทยเป็นมาตรฐานของการตัดสินเพื่อให้ได้ประโยคที่ถูกต้องเพียงประโยคเดียว

3.3 จากงานของ อติชาติ ขานทอง , วัลลภา ตันติประสงค์ชัย , ชุลีรัตน์ จรัสกุลชัย (2004) มีการนำเสนอเทคนิคและอัลกอริทึมในการสรุปใจความสำคัญของเอกสาร โดยใช้วิธีเลือกประโยคที่แสดงถึงเนื้อหาหลักของเอกสาร แล้วนำประโยคที่ได้มาสร้างเป็นใจความสำคัญของเอกสาร โดยอาศัยทฤษฎีการให้ค่าน้ำหนักของคำโดยใช้สูตรการหาค่า TF/IDF(Term Frequency / Inverse Document Frequency) รวมถึงมีการประเมินผลของใจความสำคัญที่ได้ เพื่อหาแนวทางในการสรุปใจความสำคัญของเอกสารภาษาไทย โดยในงานวิจัยนี้จะเป็นการอธิบายเฉพาะหลักในการสร้างใจความสำคัญของเอกสารซึ่งประกอบไปด้วย วัตถุประสงค์ (intent), จุดความสนใจ (focus), และขอบเขต (coverage) โดยวัตถุประสงค์ จะเป็นส่วนที่อธิบายถึงการนำใจความสำคัญไปใช้ ซึ่งแบ่งเป็น การชี้แนะ (indicative) , การให้ข้อมูล (informative) และ การประเมินค่า (evaluative) โดยเป็นการกำหนดรูปแบบของใจความสำคัญด้วยกัน 3 รูปแบบ ดังนี้

- ใจความสำคัญแบบชี้แนะ คือ ใจความสำคัญที่มีข้อมูลเพียงพอที่จะทำให้ผู้อ่านเข้าใจประเด็นหลักของเอกสารฉบับเต็ม หรือให้คำชี้แนะสั้นๆ เกี่ยวกับหัวเรื่องหลักของเอกสาร โดยทั่วไปแล้วใจความสำคัญประเภทนี้จะนำไปใช้แสดงให้แก่ผู้อ่านก่อนแสดงเอกสารฉบับเต็ม

- ใจความสำคัญแบบให้ข้อมูล คือ ใจความสำคัญที่เป็นเหมือนตัวแทนของเอกสารฉบับเต็มและจะเก็บข้อมูลรายละเอียดที่สำคัญ โดยในขณะเดียวกันก็ลดปริมาณข้อมูลที่ต้องแสดงแก่ผู้อ่านด้วย

- ใจความสำคัญแบบประเมินค่า คือ ใจความสำคัญที่จะจัดเก็บเนื้อหาตรงจุดที่เป็นผู้แต่งในหัวเรื่องที่กำหนดให้

3.4 จากงานของ ลัดดา ยินดีมาก โกรติ (2534) มีการนำเสนอเทคนิคในการค้นหาคำสำคัญเพื่อจัดทำดรรชนีสำหรับระบบการค้นคืนเอกสาร โดยอาศัยเทคนิคการให้ค่าน้ำหนักคำ คือ ความถี่การเกิดขึ้นของคำ ร่วมกับเทคนิคกำหนดจำนวนคำสำคัญ และการสร้างคำและกลุ่มคำโดยการใช้ความถี่ของ Neighborhood (Neighborhood Frequency) รวมถึงมีการกลั่นกรองคำหยุด (Stopword) และคำที่มีค่าความถี่น้อย ที่ไม่เหมาะสมที่จะนำเป็นดรรชนีในการค้นคืนด้วย

3.5 จากงานของ Hung V. Nguyen , P. Velamuru , D. Kolippakkam, H.Davulcu และ H. Lin (2003) มีการนำเสนอเทคนิคและอัลกอริทึมในการค้นหาวลีสำคัญที่ซ่อนอยู่ในเอกสารหน้าเว็บ โดยทำการค้นจากส่วนต่าง ๆ ของเอกสาร และให้น้ำหนักค่าความสำคัญในแต่ละส่วนต่าง ๆ ของเอกสารหน้าเว็บด้วย โดยนำมาสร้างให้อยู่ในรูปแบบของเมตริกซ์ค่าน้ำหนัก เช่นส่วนของ “META tags ” ส่วนของ “ Title ” ส่วนของ “ BODY tags ” โดยเมตริกซ์ค่าน้ำหนักนั้น คือ ค่าน้ำหนักความถี่การเกิดขึ้นของวลีหรือค่า TF ในแต่ละส่วนของเอกสารและค่า IDF โดยนำค่าเหล่านั้นมาสร้างเป็นเซตข้อมูลของกฎความสัมพันธ์ (Association Rule) ของคำ เพื่อใช้เทคนิคกฎความสัมพันธ์ ในการหาว่าวลีใดมีค่าน้ำหนักความสำคัญที่สุดในระดับที่กำหนดไว้

3.6 จากงานวิจัยของ Gönenç Ercan (2005) มีการนำเทคนิคการหาความสัมพันธ์ระหว่างคำภายในเอกสารเรียกว่า cohesion-based (หรือ Lexical Chains : LC method) โดยเป็นการค้นหาวลีสำคัญของเอกสารเพื่อสรุปใจความสำคัญของเอกสาร โดยเป็นการศึกษาข้อมูลในระดับผิว (surface analysis) หรือข้อมูลที่ปรากฏให้เห็น ต่อจากนั้นจึงพิจารณาดูว่าคำนามในภาษาระดับผิวนี้ มีความสัมพันธ์ทางความหมายอย่างไรกับกริยาในประโยค แล้วจึงกำหนดความสัมพันธ์ทาง

ความหมายชนิดต่างๆ เหล่านั้น โดยคุณสมบัติที่ใช้ในการพิจาณาวลีนั้น คือ ความถี่การเกิดขึ้นของวลี ตำแหน่งแรกที่เกิดขึ้นของวลีนั้น ๆ ความสัมพันธ์ระหว่างคำภายในเอกสาร และชนิดของคำสำคัญ (keyword class) รวมถึงมีการกลั่นกรองคำหยุดออกจากการค้นหาด้วย โดยทำการสร้างข้อมูลของแต่ละวลีให้อยู่ในรูปแบบของเวคเตอร์ของค่าน้ำหนักวลี และประมวลผลโดย C 4.5 Algorithm

3.7 จากงานวิจัยของ Turney (2003) นั้น ได้มีการนำจีเนติกอัลกอริทึมมาประยุกต์ ใช้ในการสกัดวลีสำคัญ โดยมีการเรียนรู้และทำการเลือกค่าที่ดีที่สุดเพื่อนำมาเป็นต้นแบบในการสกัดวลีสำคัญ โดยค่าที่ดีที่สุดของการเรียนรู้ คือโครโมโซมหรือประชากรที่มีค่าความหมาะสมเป็นประชากรต้นกำเนิด และทำการวัดค่าความเหมาะสมของแต่ละโครโมโซม เพื่อคัดเลือกเข้าสู่กระบวนจีเนติกโอเปอเรเตอร์ในขั้นตอนต่อไป โดยทำการเลือกเอาเฉพาะโครโมโซมที่มีค่าความเหมาะสมเป็นที่น่าพอใจชุดหนึ่งเก็บไว้ โดย Turney ทำการทดสอบประสิทธิภาพของแนวคิด โดยการใช้แผนภาพต้นไม้ตัดสินใจ (Decision Tree) และอัลกอริทึม C4.5 ในการทดสอบคุณสมบัติของตัวแทนเอกสาร คือ “ความถี่การเกิดขึ้นของวลี” และ “ตำแหน่งแรกที่เกิดวลีนั้น ๆ” ในเอกสารแต่ละฉบับ

การสกัดวลีสำคัญจากเอกสาร เป็นการเลือกวลีสำคัญจากคำและวลีทั้งหมดในเอกสาร วลีสำคัญที่ได้ต้องปรากฏอยู่ในเอกสารเท่านั้น โดยในปัจจุบันมีงานวิจัยด้านการค้นคืนสารสนเทศจำนวนมากที่นำโครงข่ายประสาทเทียมไปใช้ ด้วยเพราะประสิทธิภาพในด้านความแม่นยำและความสามารถในการดึงโครงสร้างออกจากข้อมูลที่มีอยู่ ซึ่งเป็นปัจจัยหลักของงานค้นคืนสารสนเทศ เช่น การใช้เทคนิคของโครงข่ายประสาทเทียมในการสกัดวลีสำคัญจากเอกสารภาษาอังกฤษ (Yair Even-Zchar : 2002)

4. วัตถุประสงค์ของงานวิจัย

4.1 เพื่อศึกษาเทคนิคการสกัดคำหรือวลีสำคัญในเอกสารภาษาไทย

4.2 เพื่อศึกษาการประยุกต์ใช้เทคนิคของโครงข่ายประสาทเทียมในการสกัดคำหรือวลี

4.3 เพื่อหาแนวทางในการเพิ่มประสิทธิภาพและปรับปรุงขั้นตอนการค้นหาของระบบการค้นคืนสารสนเทศ

5. ขอบเขตของงานวิจัย

5.1 ศึกษาเทคนิคการสกัดคำหรือวลีจากเอกสารวิชาการ

5.2 ใช้เทคนิคของโครงข่ายประสาทเทียมในการสกัดคำหรือวลี

5.3 การทดสอบความถูกต้องและการหาแนวทางการปรับปรุง

6. ขั้นตอนและวิธีดำเนินงานวิจัย

6.1 ศึกษา เทคนิคการแยกคำไทย

6.2 ศึกษา ทดสอบ และพัฒนาเทคนิคการให้น้ำหนักคำหรือวลี

6.3 ศึกษาและปรับปรุงการประมวลผลการหาใจความสำคัญ

7. ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

7.1 ได้เทคนิคที่สามารถวิเคราะห์และเพิ่มประสิทธิภาพของผลการค้นหาข้อมูลในเอกสาร รวมถึงเทคนิคสำหรับสกัดคำหรือวลีที่มีความเหมาะสมและสามารถเป็นตัวแทนของเอกสารได้ดี

7.2 ได้แนวทางนำที่ได้จากข้อ 7.1 มาใช้เป็นเทคนิคในการดำเนินการและปรับปรุงผลการค้นหาของระบบการค้นคืน