รายงานการวิจัย

การวิเคราะห์ข้อความเว็บเอกสาร

Web Content Analysis

ผู้วิจัย ผศ.ดร. พรฤดี เนติโสภากุล

ได้รับทุนสนับสนุนงานวิจัยจากเงินรายได้ประจำปีงบประมาณ 2552

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

กิตติกรรมประกาศ

งานวิจัยเรื่อง การวิเคราะห์เนื้อความเว็บเอกสาร ผู้วิจัยได้ทำการเปรียบเทียบงานวิจัยที่ผ่านมาและนำเสนอแนวทางการวิเคราะห์เนื้อความเว็บเอกสารโดยการใช้ออนโทโลจีเป็นฐานความรู้สำหรับระบบ ทำงานร่วมกับเทคนิคในการประมวลผลภาษาธรรมชาติ ซึ่งงานวิจัยนี้ได้รับทุนจากคณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง จึงขอกราบขอบพระคุณเป็นอย่างสูงมา ณ ที่นี้

พรฤดี เนติโสภากุล

บทคัดย่อ

ชื่อโครงการ (ภาษาไทย) การวิเคราะห์ข้อความบนเว็บเอกสาร

(ภาษาอังกฤษ) Web Content Analysis

ได้รับทุนอุดหนุนการวิจัยจาก คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ประจำปี 2552 จำนวนเงิน 20,000 บาท

ระยะเวลาทำการวิจัย ปี ตั้งแต่ 1 ตุลาคม พ.ศ. 2551 ถึง 30 กันยายน พ.ศ. 2552

ผู้ดำเนินการวิจัย ผศ.ดร. พรฤดี เนติโสภากุล คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง เบอร์โทรศัพท์ 02-723-4957

บทคัดย่อ (ภาษาไทย)

งานวิจัยนี้ศึกษาและทำการเปรียบเทียบแนวทางการออกแบบและการวิเคราะห์ข้อความเว็บเอกสารด้วยเทคนิคต่าง ๆ พร้อมทั้งนำเสนอแนวทางการวิเคราะห์ข้อความเว็บเอกสาร โดยการนำออนโทโลจีในการแทนความรู้ให้กับระบบ และทำงานร่วมกับการประมวลผลภาษาธรรมชาติ โดยผลลัพธ์จากการวิเคราะห์ความหมายนี้ จะแสดงในรูปแบบของเฟรม สล็อต ซึ่งสามารถนำไปใช้ประโยชน์เป็นฐานความรู้ให้กับแอพพลิเคชั่นอื่นๆ ได้ เช่น ใช้เป็นฐานความรู้ในระบบถามตอบต่อไป

บทคัดย่อ (ภาษาอังกฤษ)

This research studies and compares web content analysis techniques. Then presents an approach to analyze web content by representing a system knowledge using ontology, working together with natural language processing techniques. The results of web content analysis will be stored in a frame-slot format, which can be used as a knowledge base for other applications such as a question-answering system.

สารบัญ

สารบัญตาราง

สารบัญรูป

บทที่ 1
บทนำ

1.1 ปัญหาและความเป็นมา

ในปัจจุบันอินเทอร์เน็ตเป็นแหล่งความรู้ขนาดใหญ่ที่ประกอบด้วยข่าวสารข้อมูลที่มีความหลากหลายและมีรูปแบบในการนำเสนอที่แตกต่างกัน เช่น ข้อความ ตาราง รูปภาพ เป็นต้น แต่อย่างไรก็ตามข่าวสารโดยส่วนใหญ่จะอยู่ในรูปแบบของข้อความ ซึ่งมีลักษณะที่ไม่เป็นโครงสร้าง โดยที่ข้อความเหล่านั้นเป็นรูปแบบที่มนุษย์สามารถเข้าใจความหมายของเนื้อหาได้ แต่งานวิจัยด้านการทำความเข้าใจเนื้อหา ยังเป็นงานที่น่าท้าทายอย่างยิ่ง ในงานวิจัยนี้จะศึกษาเรื่องการวิเคราะห์ข้อความจากเว็บเอกสาร ซึ่งมุ่งหมายให้คอมพิวเตอร์สามารถสกัดข้อความที่สำคัญจากเอกสารได้แบบอัตโนมัติ โดยที่ข้อความได้จากสกัดได้จะเก็บในรูปแบบที่เป็นโครงสร้างที่คอมพิวเตอร์สามารถนำผลลัพธ์นั้นไปประมวลผลต่อได้โดยง่าย เช่น ในระบบถามตอบ (Question-Answering)

งานวิจัยด้านการวิเคราะห์ข้อความจากเว็บเอกสารนั้นสามารถแบ่งออกได้เป็น 2 กลุ่มคือ (1) การวิเคราะห์เอกสารเฉพาะโดเมน (Domain Specific) (Lyons and Smith, 2002, Tellez-Valero et al., 2009, Alani et al., 2003, Mestrovic et al., 2007, Laclavik et al.,2009) เช่น จากเอกสารสัมมนา จากข่าวการพยากรณ์อากาศ ข่าวการเกษตร ข่าวกีฬา เป็นต้น พบว่าข้อความในแต่ละกลุ่มนั้น ก็มีความแตกต่างกันของเนื้อหาอย่างชัดเจนและสิ่งที่ต้องการวิเคราะห์ก็มีความแตกด้วยเช่นกัน (2) การวิเคราะห์ข้อความที่สามารถทำงานได้กับเอกสารทุกเรื่อง (Open Domain) (Popv et al., 2003, Moschitti et al., 2003, Banko et al., 2007, Pasca 2009) จะเป็นการวิเคราะห์ข้อความในโดเมนทั่ว ๆ ไป และข้อความที่ต้องการวิเคราะห์นั้นจะไม่มีลักษณะเฉพาะที่ชัดเจนเหมือนกับการวิเคราะห์ข้อความที่ขึ้นกับโดเมนของเนื้อหาเฉพาะด้าน

กระบวนการวิเคราะห์ข้อความจะอาศัยเทคนิคการประมวลผลภาษาธรรมชาติ ในการวิเคราะห์โครงสร้างไวยากรณ์ของข้อมูล เทคนิคทางสถิติ โดยการเรียนรู้โดยอัตโนมัติจากคลังข้อความที่ได้มีการกำกับแท็คที่ต้องการ และในปัจจุบันได้มีการใช้ออนโทโลจีเข้ามามีบทบาทในการวิเคราะห์ข้อความ โดยทำให้คอมพิวเตอร์สามารถเข้าใจความหมายของกลุ่มคำในเอกสาร และแสดงความสัมพันธ์ของข้อมูลที่เกี่ยวข้องได้ เช่น การสกัดข้อมูลส่วนบุคคล ออนโทโลจีสามารถแสดงความสัมพันธ์ที่เกี่ยวข้องกับบุคคลนั้นได้ เช่น สถานที่ทำงาน ตำแหน่งงาน เป็นต้น

ดังนั้น ในงานวิจัยนี้จึงได้ศึกษาเปรียบเทียบแนวทางการออกแบบและการวิเคราะห์ข้อความเว็บเอกสารด้วยเทคนิคต่าง ๆ และนำเสนอแนวทางการวิเคราะห์ข้อความเว็บเอกสาร โดยการนำออนโทโลจี มาใช้งานร่วมกับการประมวลผลภาษาธรรมชาติ โดยผลลัพธ์จากการวิเคราะห์

ความหมาย จะแสดงอยู่ในรูปแบบของเฟรม สล็อต (Frame-Slot) ที่มีการจัดเก็บข้อมูลอย่างเป็นโครงสร้าง และข้อมูลในแต่ละสล็อตของเฟรมจะแสดงถึงองค์ประกอบต่าง ๆ ที่สนใจในโดเมนนั้น ๆ

1.2 วัตถุประสงค์ในการวิจัย

1. เพื่อศึกษางานวิจัยทางด้านการวิเคราะห์ข้อความเว็บเอกสาร โดยมีการเปรียบเทียบข้อดีและข้อเสียในแต่ละงานวิจัย
2. ศึกษาการออกแบบเครื่องมือในการวิเคราะห์ข้อความเว็บเอกสาร

1.3 ขอบเขตการศึกษา

ในการวิจัยนี้แสดงลำดับขั้นตอนการศึกษา แสดงดังรายละเอียดต่อไปนี้

1. ศึกษาทฤษฎี และงานวิจัยจากบทความ และเอกสารต่าง ๆ ที่เกี่ยวข้องกับงานวิจัยนี้
2. ศึกษาสถาปัตยกรรมและเปรียบเทียบแต่ละเทคนิคการวิเคราะห์ข้อความเว็บเอกสาร
3. นำเสนอแนวทางการออกแบบการวิเคราะห์ข้อความเว็บเอกสาร โดยข้อมูลที่ใช้ในการวิเคราะห์ข้อความข่าวในโดเมนสภาพอากาศ (Weather News)
4. สรุปและเขียนรายงาน
5. นำเสนอผลงานวิจัยในรูปแบบบทความงานประชุม

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อให้ทราบข้อแตกต่างในแต่ละแนวทางการออกแบบการวิเคราะห์ข้อความเว็บเอกสาร และสามารถเป็นแนวทางในการศึกษาในอนาคต
2. สามารถพัฒนาต่อยอดงานทางด้านการวิเคราะห์ข้อความเว็บเอกสาร

บทที่ 2
แนวคิดและเทคนิคที่เกี่ยวข้อง

ในการศึกษาวิจัยนี้มุ่งเน้นการวิเคราะห์เปรียบเทียบเทคนิคด้านการวิเคราะห์ข้อความจากเว็บเอกสารแบบอัตโนมัติ โดยผู้วิจัยได้ศึกษาเอกสาร เทคนิค ทฤษฎี และงานวิจัยที่เกี่ยวข้อง โดยดำเนินการศึกษาใน 2 ประเด็น ดังนี้

1. แนวคิดการวิเคราะห์ข้อความ
2. เทคนิคที่ใช้ในการวิเคราะห์ข้อความจากเอกสารแบบอัตโนมัติ โดยแบ่งออกเป็น 4 กลุ่ม ได้แก่ การวิเคราะห์ข้อความด้วยวิธีการเปรียบเทียบรูปแบบ (Pattern-Matching) การวิเคราะห์ข้อความด้วยวิธีการเรียนรู้ของเครื่องจักร (Machine Learning) การวิเคราะห์ข้อความโดยการใช้ออนโทโลจี และการวิเคราะห์โดยการใช้ฐานความรู้อื่น

2.1 การวิเคราะห์ข้อความ (Content Analysis)

การวิเคราะห์เนื้อความเอกสาร เป็นเทคนิคเพื่อทำการสรุปเป้าหมายและแยกลักษณะเฉพาะของข้อความในเอกสารอย่างเป็นระบบ (Holsti, 1969)

การวิเคราะห์ข้อความ เป็นการสรุป วิเคราะห์ข้อความที่มีจำนวนมาก โดยใช้วิธีการทางวิทยาศาสตร์ ที่มีการระบุวัตถุประสงค์ เครื่องมือที่ใช้ วิธีการออกแบบการวิเคราะห์ มีความน่าเชื่อถือ และสามารถตรวจสอบได้ (Kimberly, 2002)

การวิเคราะห์ข้อความเอกสารเป็นการนำข้อมูลมาจัดให้เป็นโครงสร้างที่เป็นระเบียบ โดยศึกษาความความหมายและความสัมพันธ์ของข้อความที่ปรากฏในเอกสาร โดยเป็นการระบุว่า ใคร ทำอะไร ที่ไหน เมื่อไหร่ อย่างไร เพื่อให้เกิดความชัดเจนในข้อมูล โดยทำให้ทราบขอบเขต และรายละเอียดของเนื้อหาอย่างละเอียด แล้วนำไปสู่การเกิดความรู้ใหม่ เช่น ข้อมูลข่าวหนังสือพิมพ์ในประเด็นข่าวสภาพอากาศ ซึ่งเราต้องการทราบว่า เกิดอากาศลักษณะอย่างไร สถานที่ วันเวลา มีผลกระทบต่อใคร อย่างไร โดยคำตอบเหล่านี้ก็มาจากการวิเคราะห์ข้อความในเอกสาร

ข้อความที่ใช้ในการวิเคราะห์ ได้แก่ เอกสาร บทสัมภาษณ์หรือบทสนทนา ข่าวหนังสือพิมพ์ โฆษณา บทความ เว็บเอกสาร เป็นต้นโดยเริ่มต้นนั้นงานทางด้านการวิเคราะห์จะทำการวิเคราะห์เอกสารหนังสือพิมพ์ซึ่งอยู่ในรูปแบบเล่มเอกสารที่เป็นกระดาษ ต่อมาจึงใช้กับการวิเคราะห์ข้อมูลโฆษณา บทความ บทสนทนา แต่ในปัจจุบันข้อมูลที่นิยมใช้ในการวิเคราะห์คือข้อมูลบนอินเทอร์เน็ตซึ่งมีอยู่เป็นจำนวนมากและมีความหลากหลาย อีกทั้งยังเป็นแหล่งข้อมูลที่สามารถเข้าถึงได้ง่าย โดยภายในข้อมูลเหล่านั้นได้ปรากฏความรู้ซ่อนอยู่ และเมื่อทำการวิเคราะห์ข้อมูลเหล่านั้นจะได้ความรู้ที่สำคัญที่สามารถนำไปใช้ประโยชน์ต่อได้ ทำให้เกิดงานวิจัยต่าง ๆ ที่

พยายามตอบคำถามที่ว่า เราจะนำความรู้เหล่านั้นออกมาได้อย่างไร ด้วยวิธีอะไร ซึ่งในยุคแรกได้มีการวิเคราะห์และนำความรู้เหล่านั้นมาใช้ได้โดยการใช้เครื่องมือสืบค้นในการเลือกเอกสารให้สอดคล้องกับคำสำคัญที่ใช้ในการสืบค้น ทำให้สามารถได้เอกสารที่ผ่านการคัดกรองแล้วว่าเกี่ยวข้องกับคำสำคัญที่ต้องการสืบค้น แต่อย่างไรก็ตามข้อความที่ต้องการอาจประกอบอยู่ในทั้งเอกสารหรือปรากฏเป็นส่วนหนึ่งของเอกสาร ซึ่งยังต้องการทำวิเคราะห์ต่อไป โดยการใช้เทคนิคการสกัดความรู้จากเอกสาร ซึ่งในแต่ละโดเมนจะมีการกำหนดล่วงหน้าก่อนว่าต้องการสกัดข้อมูลอะไร แล้วทำการวิเคราะห์ข้อความในเอกสารเพื่อให้ได้คำตอบตามข้อมูลที่ต้องการ

โดยผลลัพธ์ที่ได้จากการวิเคราะห์นั้นจะต้องง่ายต่อการเข้าใจและสามารถนำไปใช้ประโยชน์ต่อในอนาคตได้ สิ่งที่ต้องการจากการวิเคราะห์ข้อความคือ เพื่อสรุปข้อมูล เพื่อหารูปแบบหรือความสัมพันธ์ภายในเอกสาร และเพื่อหารูปแบบหรือความสัมพันธ์จากภายนอกเอกสาร

2.2 เทคนิคการวิเคราะห์ข้อความ

เทคนิคในการวิเคราะห์ข้อความเอกสารอัตโนมัติ ในงานวิจัยนี้จะแบ่งออกเป็น 4 กลุ่ม ได้แก่ การวิเคราะห์ข้อความด้วยวิธีการเปรียบเทียบรูปแบบ (Pattern-Matching) การวิเคราะห์ข้อความด้วยวิธีการเรียนรู้ของเครื่องจักร (Machine Learning) การวิเคราะห์ข้อความโดยการใช้ออนโทโลจี และการวิเคราะห์โดยการใช้ฐานความรู้อื่น ซึ่งในการวิเคราะห์ข้อความนี้ในงานวิจัยโดยส่วนใหญ่นิยมทำงานกับข้อมูลเฉพาะเรื่อง เนื่องจากว่าแต่ละข้อความในเอกสารเฉพาะเรื่องก็มีความแตกต่างกันในด้านของความหมาย ความสัมพันธ์ภายในเอกสาร และผลลัพธ์ที่ต้องการจากการวิเคราะห์นั้นก็มีความแตกต่างกัน โดยแยกอธิบายแต่ละเทคนิคดังนี้

2.2.1 เทคนิคการวิเคราะห์ข้อความด้วยวิธีการเปรียบเทียบรูปแบบ (Pattern-Matching)

การเปรียบเทียบรูปแบบเป็นวิธีการที่นำแพทเทิร์นที่ได้กำหนดไว้ในรูปแบบต่าง ๆ นั้นเปรียบเทียบกับข้อความในเอกสารที่กำหนด ซึ่งรูปแบบของแพทเทิร์นที่กำหนดนั้นสามารถกำหนดได้ในหลากหลายรูปแบบ เช่น กำหนดแพทเทิร์นตามรูปแบบโครงสร้างไวยากรณ์ของประโยคเพื่อสกัดข้อมูลที่ต้องการ (Rilloff and Schelzanbach, 1998)

Riloff and Schelzenbach (1998) ได้เสนองานวิจัยในการสกัดข้อมูลที่ต้องการจากเอกสารแล้วแล้วจัดเก็บในรูปแบบของเฟรมสล็อต โดยเป็นการสร้างเฟรมที่กำหนดข้อมูลที่ต้องการสกัดและข้อมูลที่ต้องการสกัดจะกำหนดในแต่ละสล็อต ซึ่งระบบที่พัฒนาขึ้นนั้นมีชื่อว่า AutoSlog-TS ในการกำหนดข้อมูลแต่ละเฟรมนั้น ระบบจะทำการสร้างแพทเทิร์นต่าง ๆ ที่ใช้ในการสกัดข้อมูลจากเอกสารในโดเมนข่าวก่อการร้ายในลาตินอเมริกาแล้วสกัดข้อมูลตามแพทเทิร์นที่ระบุ

การกำหนดเฟรมที่ใช้ในการสกัดข้อมูลจะระบุสล็อตที่ต้องการด้วย เช่น

Frame: active –verb denonated

Slot: perpetrator	subject	TERRORIST
Slot: instrument	subject	WEAPON
Slot: instrument	direct-obj	WEAPON

จากประโยค The guerrillas detonated a bomb. ซึ่งสอดคล้องกับเฟรม active –verb denonated จึงสกัดข้อมูลใส่ในแต่ละสล็อตได้ดังนี้

Frame: active –verb detonated

Slot: perpetrator	The guerrillas
Slot: instrument	bomb

ซึ่งมีขั้นตอนในงานวิจัยนี้ดังต่อไปนี้

1. การสร้างแพทเทิร์นจากเอกสารกลุ่มตัวอย่าง เพื่อระบุข้อความที่อยู่รอบข้างข้อความที่เป็นเป้าหมาย โดยวิเคราะห์โครงสร้างทางไวยากรณ์ของประโยคเพื่อระบุนามวลี แล้วสร้างแพทเทิร์นขึ้นมาโดยเทียบกับกฎที่กำหนด เช่น กฎ <subject>active-verb, active-verb<dobj> เป็นต้น จากนั้นนำแพทเทิร์นที่ได้มาหาค่าทางสถิติ เพื่อดูอัตราการเกิดขึ้นของแพทเทิร์นที่ถูกเรียกใช้ในแต่ละประโยค เรียงลำดับแพทเทิร์นเพื่อดูความเกี่ยวข้องกับเอกสาร ตัวอย่างแพทเทิร์นที่สกัดได้ <subj> detonated, <subject> was killed, <subject> was kidnapped, claim<dobj> เป็นต้น โดยแพทเทิร์นที่ได้นี้จะใช้เทียบกับประโยคเพื่อสกัดข้อมูลที่ต้องการจากเอกสาร

2. สร้าง semantic lexicon (พจนานุกรมคำศัพท์ที่มีการจัดหมวดหมู่ตามความหมาย) ของแต่ละ Semantic Category ได้แก่ TERRORIST, LOCATION, WEAPON, TIME เป็นต้น โดยกำหนดค่าเริ่มต้นของ seed word ในแต่ละกลุ่มด้วยคำจำนวน 5 คำ หานามหลักจากนามวลีในแต่ละประโยค เลือกประโยคที่เป็น seed word จากนั้นหาคำนามที่อยู่ติดกันด้านหน้าและด้านหลังของ seed word ว่าคือคำอะไร เพื่อจัดกลุ่มของคำนามซึ่งคิดว่าเป็นกลุ่มเดียวกันกับ seed word และมีการคำนวณคะแนนการจัดกลุ่มของแต่ละคำโดยคำนวณจาก ความถี่ของคำในกลุ่มในแต่ละประโยค/ความถี่ของคำใน corpus จากนั้นเลือกคำที่อยู่ในกลุ่มของ stopword และคำที่มีความถี่น้อยกว่าเท่ากับ 5 ออกจากกลุ่มที่สนใจ เรียงลำดับคำโดยแยกตามคะแนนของกลุ่ม เลือกคำมา 5 คำจากอันดับสูงสุดเก็บไว้ใน seed word lists

3. สร้าง semantic profile ของแต่ละแพทเทิร์น โดยนำแพทเทิร์นมาวิเคราะห์กับเอกสารว่าประกอบด้วยคำนามอะไร และคำนามนั้นแต่ละคำจัดอยู่ในกลุ่มใด โดยดูจาก semantic lexicon จะได้ semantic category จากนั้นคำนวณหาค่า PFreq (จำนวนครั้งของการปรากฏ pattern ในโดเมน) SFreq (จำนวนครั้งของการปรากฏคำในแต่ละกลุ่ม) และค่า Prob = SFreq/PFreq แล้วทำการเลือก semantic preference จาก semantic profile โดยมีเงื่อนไขคือ (SFreq >= F1) or ((SFreq >=

F2) and (Prob >= P)) ซึ่ง semantic profile จะเป็นการกำหนดว่าแต่ละแพทเทิร์นนั้นควรประกอบด้วย semantic category อะไร

4. สร้าง conceptual roles ของแต่ละแพทเทิร์น โดยกำหนดตาม domain role ซึ่งก็คือชื่อสล็อตกำหนดไว้ล่วงหน้าแล้ว และแต่ละแพทเทิร์นจะถูกขยายรายละเอียดเพิ่มเติมโดยเพิ่มลงใน conceptual role ตัวอย่างเช่น

แพทเทิร์น <subj> detonated

perpetrator TERRORIST

instrument WEAPON

โดยที่ perpetrator และ instrument คือ Domain Role และ TERRORIST และ WEAPON คือ Semantic Category

5. กำหนดเฟรมตามแพทเทิร์นที่กำหนด แล้วระบุสล็อตรายละเอียดลักษณะของข้อมูลที่ต้องการสกัดของแต่ละสล็อตของเฟรม

การทำงานของระบบ AutoSlog-TS จะใช้แพทเทิร์นเป็นหลักซึ่งในงานวิจัยนี้ยังคงต้องนำแพทเทิร์นที่สกัดจากระบบมาผ่านการพิจารณาจากผู้เชี่ยวชาญอีกครั้ง ว่าควรใช้แพทเทิร์นใดบ้างจึงจะเหมาะสมกับโดเมน ถึงแม้ว่าในขั้นตอนของการสร้างแพทเทิร์นนั้นจะสามารถทำได้แบบอัตโนมัติแล้วก็ตาม แต่เพื่อความถูกต้องในการสกัดข้อมูลนั้นผู้เชี่ยวชาญก็ยังจำเป็นอยู่

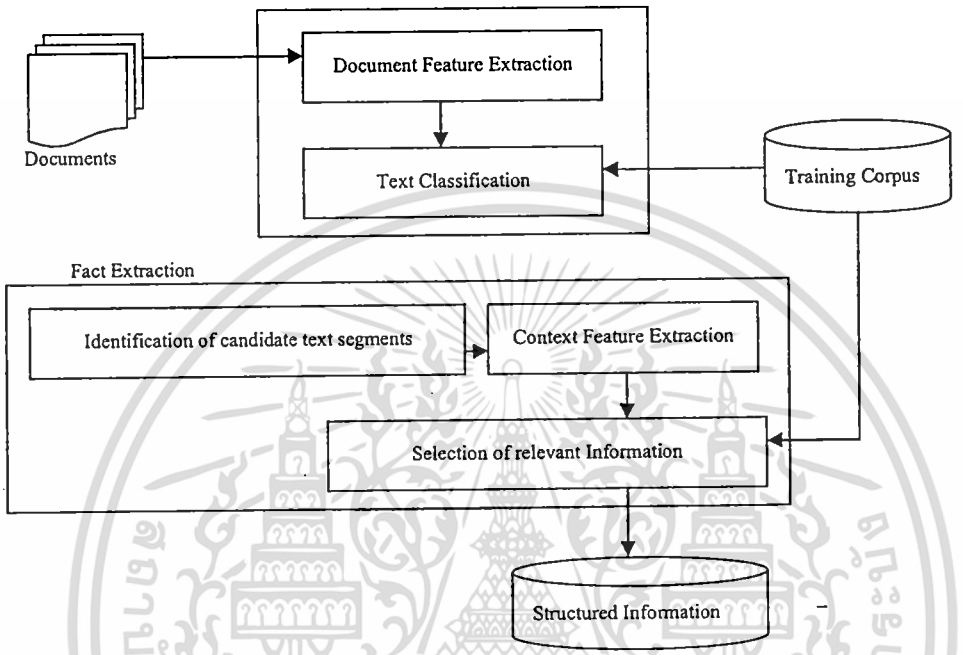
2.2.2 เทคนิคการวิเคราะห์ข้อความเอกสารด้วยวิธีการเรียนรู้ด้วยเครื่องจักร

การเรียนรู้ด้วยเครื่องจักร เป็นการทำให้เครื่องเรียนรู้จากข้อมูลตัวอย่าง และความรู้ที่เรียนรู้ได้จะถูกจัดเก็บในฐานความรู้ที่กำหนด เพื่อใช้ทำการวิเคราะห์กับข้อมูลใหม่ โดยเรียนรู้จากข้อมูลเดิม ซึ่งความรู้ที่ได้จากการเรียนรู้ด้วยเครื่องจักรนี้ขึ้นอยู่กับจำนวนข้อมูลตัวอย่างซึ่งจะต้องมีขนาดใหญ่ เพื่อให้เครื่องสามารถเรียนรู้เนื้อหาของโดเมนต่าง ๆ ได้อย่างครอบคลุม และอัลกอริทึมที่ใช้ในการเรียนรู้ซึ่งประกอบด้วยหลายเทคนิคด้วยกัน เช่น Support Vector Machine, Decision Tree, Naïve Bayes เป็นต้น

Tellez-Valero, et al. (2009) ได้ทำการสกัดข้อความที่สำคัญจากข่าวออนไลน์เกี่ยวกับภัยธรรมชาติแบบอัตโนมัติด้วยวิธีการเรียนรู้ด้วยเครื่องจักร (Machine Learning) โดยเป็นการให้เครื่องเรียนรู้จากข้อมูลตัวอย่างที่กำหนดไว้ แล้วเมื่อพบข้อมูลใหม่เครื่องก็จะทำการเรียนรู้ข้อมูลใหม่นั้นจากข้อมูลที่มีอยู่ด้วยอัลกอริทึมต่าง ๆ

การสกัดข้อมูลในงานวิจัยนี้มุ่งเน้นการทำงานในการสกัดข้อมูลเกี่ยวกับภัยธรรมชาติ ได้แก่ ภัยธรรมชาติจากเฮอริเคน ไฟไหม้ป่า ภาวะแห้งแล้ง แผ่นดินไหว และน้ำท่วม โดยข้อมูลที่ต้องการสกัดจากเอกสารประกอบด้วย วันที่เกิดภัยธรรมชาติ สถานที่เกิดเหตุ จำนวนผู้เสียชีวิต จำนวนผู้ได้รับบาดเจ็บ จำนวนผู้ได้รับความเสียหายจากภัยธรรมชาติ จำนวนผู้ได้รับผลกระทบจากภัย

ธรรมชาติ จำนวนผู้สูญหาย จำนวนที่อยู่อาศัยที่ถูกทำลาย จำนวนที่อยู่อาศัยที่ได้รับผลกระทบ จำนวนพื้นที่ได้รับผลกระทบ ความเสียหายทางเศรษฐกิจ และข้อมูลที่สกัดเหล่านี้จะจัดเก็บในรูปแบบของเฟรม

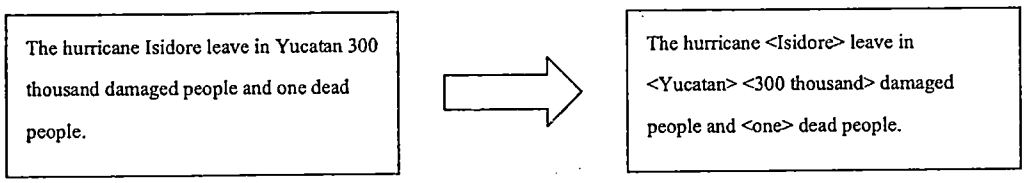


ภาพที่ 2.1 แสดงสถาปัตยกรรมของระบบในงานวิจัย Tellez-Valero, et al. (2009)

จากภาพที่ 2.1 แสดงขั้นตอนการสกัดข้อมูลในงานวิจัยนี้ โดยเริ่มต้นจะทำการคัดเลือกเอกสารก่อนว่าเอกสารนั้นเป็นเอกสารที่เกี่ยวข้องกับข้อมูลที่ต้องการสกัดหรือไม่ เพื่อแยกเอกสารที่ไม่เกี่ยวข้องออกโดยในงานวิจัยนี้ได้ทดลองการแยกกลุ่มเอกสารด้วยเทคนิคจำนวน 3 เทคนิค ได้แก่ Support Vector Machine (SVM), Naïve Bayes และ C4.5 เพื่อหาเทคนิคที่ดีที่สุดในการแยกกลุ่มเอกสารของงานวิจัยนี้ จากการทดลองพบว่าการจัดกลุ่มเอกสารด้วยวิธี SVM นั้นให้ประสิทธิภาพที่ดีกว่าอีกสองวิธี

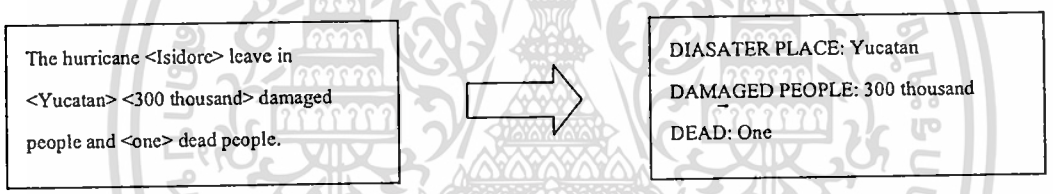
เมื่อระบบทำการเลือกเอกสารที่เกี่ยวข้องแล้ว จากนั้นทำการสกัดข้อมูลที่ต้องการออกจากเอกสารนั้น ประกอบด้วย 2 ขั้นตอนคือ

1. ระบุข้อมูลที่มีความเป็นไปได้ที่จะเป็นข้อมูลที่ต้องการสกัดซึ่งกำหนดไว้ล่วงหน้าแล้วว่าต้องการสกัดข้อมูลอะไรบ้าง โดยใช้นิพจน์ระบุนาม (Regular Expression) ในการวิเคราะห์ข้อมูลที่ต้องการ ซึ่งข้อมูลที่วิเคราะห์นั้นจะเป็นข้อมูลจำพวก ชื่อเฉพาะ จำนวนตัวเลข และวันที่ แสดงดังภาพที่ 2.2



ภาพที่ 2.2 แสดงการระบุข้อมูลที่เกี่ยวข้องที่ต้องการสกัดในเอกสาร โดยข้อมูลที่มีความเป็นไปได้นั้น จะอยู่ภายใต้เครื่องหมาย <>

2. เลือกข้อมูลที่เกี่ยวข้องที่จะเป็นผลลัพธ์ของการสกัดข้อมูลจากเอกสาร เพื่อนำไปใส่ในเฟรม โดยในการเลือกข้อมูลว่าควรจะเกี่ยวข้องกับหัวข้อใดนั้นใช้เทคนิคการแบ่งกลุ่มเอกสาร โดยแบ่งเป็นกลุ่มข้อมูลเกี่ยวกับชื่อเฉพาะ จำนวนตัวเลข และวันที่ และ feature ที่ใช้ในการแบ่งกลุ่มข้อมูลนั้นจะอาศัยบริบทข้างเคียงในเอกสาร จากภาพที่ 2.3 แสดงการเลือกข้อความมาใส่ในเฟรม พบว่า คำว่า "Isidore" จะไม่ถูกเลือกนำมาใส่ในเฟรม เนื่องจากชื่อเฮอร์ริเคนไม่ใช่ข้อมูลที่ต้องการสกัด ข้อมูลจะถูกเลือกนำมาใส่ในแต่ละสล็อต แสดงดังภาพที่ 2.3



ภาพที่ 2.3 แสดงข้อมูลที่สกัดจากเอกสารแล้วแสดงในรูปแบบของเฟรม

จากการทดลองในงานวิจัยนี้ไม่มีการใช้เทคนิคของการประมวลผลภาษาธรรมชาติ นั้นคือไม่ได้วิเคราะห์โครงสร้างและความหมายของประโยค เพียงแต่ให้เครื่องจักรเรียนรู้จากข้อมูลตัวอย่างแล้วทำการวิเคราะห์ข้อมูลใหม่ว่าข้อมูลอะไรภายในเอกสารที่ต้องการสกัด โดยส่วนใหญ่แล้วข้อมูลที่ต้องการสกัดนี้เป็นเพียงข้อมูลที่เป็นชื่อเฉพาะ ตัวเลขและวันที่ จึงไม่จำเป็นต้องอาศัยเทคนิคของการประมวลผลภาษาธรรมชาติก็ได้ แต่ถ้าข้อมูลที่ต้องการสกัดมีความซับซ้อนมากขึ้น การใช้เทคนิคการเรียนรู้ของเครื่องจักรเพียงอย่างเดียวนั้น อาจจะทำให้ความถูกต้องของข้อมูลที่ต้องการสกัดมีประสิทธิภาพไม่ดีเท่าที่ควร

2.2.3 เทคนิคการวิเคราะห์ข้อความโดยการใช้ออนโทโลจี

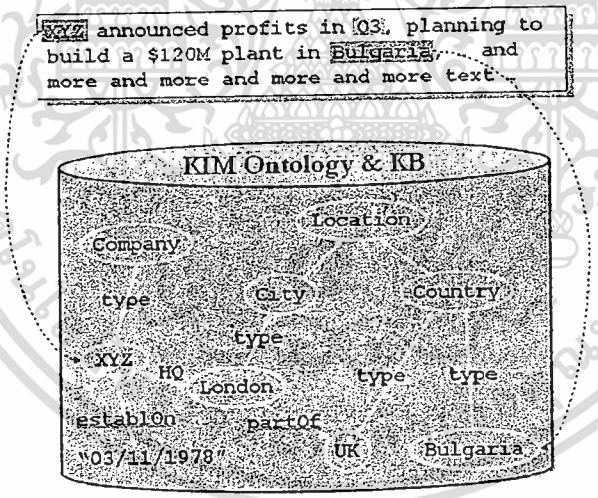
ออนโทโลจีและการสกัดข้อมูลมีความเกี่ยวข้องกัน 2 ด้าน คือ

1. ออนโทโลจีจะถูกใช้เป็นส่วนหนึ่งในขั้นตอนการสกัดข้อมูล โดยที่ออนโทโลจีเป็นส่วนหนึ่งในการทำให้คอมพิวเตอร์สามารถเข้าใจความหมายของการสกัดข้อมูลที่ต้องการ

2. ในการเพิ่มอินสแตนซ์ให้กับออนโทโลจี ซึ่งเป็นการขยายออนโทโลจีให้มีขนาดใหญ่ขึ้นนั้นต้องอาศัยการสกัดข้อมูลในการสกัดอินสแตนซ์จากข้อความ

จะพบว่าทั้งออนโทโลจีและการสกัดข้อมูลเป็นงานที่สามารถทำงานร่วมกันโดยออนโทโลจีถูกใช้สำหรับการอธิบายความหมายของข้อความเพื่อให้การสกัดข้อมูลมีประสิทธิภาพมากยิ่งขึ้น และในการสกัดข้อมูลนั้นสิ่งที่สกัดได้คือความรู้ใหม่ที่เมื่อเราได้เพิ่มเข้าไปเป็นส่วนหนึ่งในออนโทโลจีก็จะทำให้ออนโทโลจีนั้นมีความครอบคลุมมากยิ่งขึ้น

Popov, et al. (2003) ได้นำเสนอเครื่องมือในการกำกับความหมายให้กับเว็บเอกสารแบบอัตโนมัติ เรียกว่า KIM ซึ่งได้มีการใช้เทคนิคการสกัดข้อมูลมาทำงานร่วมกับออนโทโลจี โดยออนโทโลจีจะทำหน้าที่ในการเก็บรวบรวมข้อมูลและความสัมพันธ์ระหว่างข้อมูลที่เกี่ยวข้องที่ต้องการกำกับความหมาย และใช้เทคนิคการสกัดข้อมูลในการสกัดข้อความในเอกสารเพื่อทำการกำกับความหมายตามคอนเซ็ปในออนโทโลจี จากภาพที่ 2.4 คำในเอกสารจะถูกกำกับด้วยคอนเซ็ปในออนโทโลจีที่คำนั้นสอดคล้องกับอินสแตนซ์ในออนโทโลจี แต่อย่างไรก็ตามการกำกับความหมายจะไม่ได้กำกับลงไปที่เอกสาร ภาพที่ 2.5 แสดงตัวอย่างเอกสารที่ถูกกำกับความหมาย



ภาพที่ 2.4 ตัวอย่างการกำกับความหมายให้กับข้อความ

นอกจากที่จะแสดงคอนเซ็ปของคำในเอกสารแล้ว ระบบจะแสดงความสัมพันธ์ คุณสมบัติอื่น ๆ และเอกสารที่เกี่ยวข้องกับคำนั้นที่ถูกกำกับความหมาย เช่น คอนเซ็ปของ United States คือ Country แสดงความสัมพันธ์ในออนโทโลจีได้ดังนี้

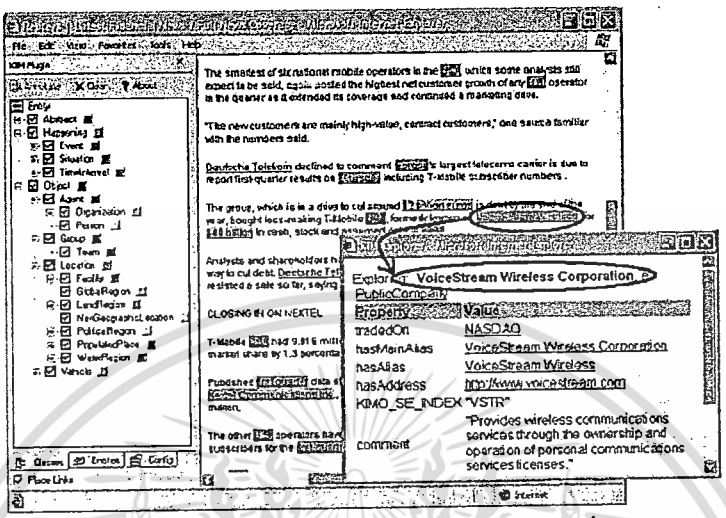
United States *hasAlias* USA

United States *hasAlias* U.S.A

United States *hasAlias* United State of America

United States *hasCapital* Washington, D.C.

United States *locatedIn* North America



ภาพที่ 2.5 แสดงตัวอย่างเอกสารที่กำกับความหมายโดยแยกตามคอนเซ็ปที่ปรากฏในออนโทโลจี

ในการกำกับความหมายนั้น คอนเซ็ปที่ใช้ในการกำกับจะพิจารณาจากออนโทโลจีที่ได้สร้างขึ้นไว้ก่อนแล้ว เรียกว่า KIM Ontology ซึ่งประกอบด้วย 250 คอนเซ็ปและ 100 ความสัมพันธ์ ส่วนอินสแตนซ์จะจัดเก็บไว้ใน KIM Knowledge Base (KIM KB) โดยอินสแตนซ์ที่จัดเก็บจะเป็นชื่อสถานที่ ชื่อบุคคล ชื่อหน่วยงาน เป็นต้น ซึ่งอินสแตนซ์นั้นสามารถเรียนรู้เพิ่มขึ้นได้นอกจากที่จะกำหนดไว้ในออนโทโลจี โดยใช้เทคนิคของการสกัดข้อมูลในการสกัดอินสแตนซ์ของแต่ละคอนเซ็ป

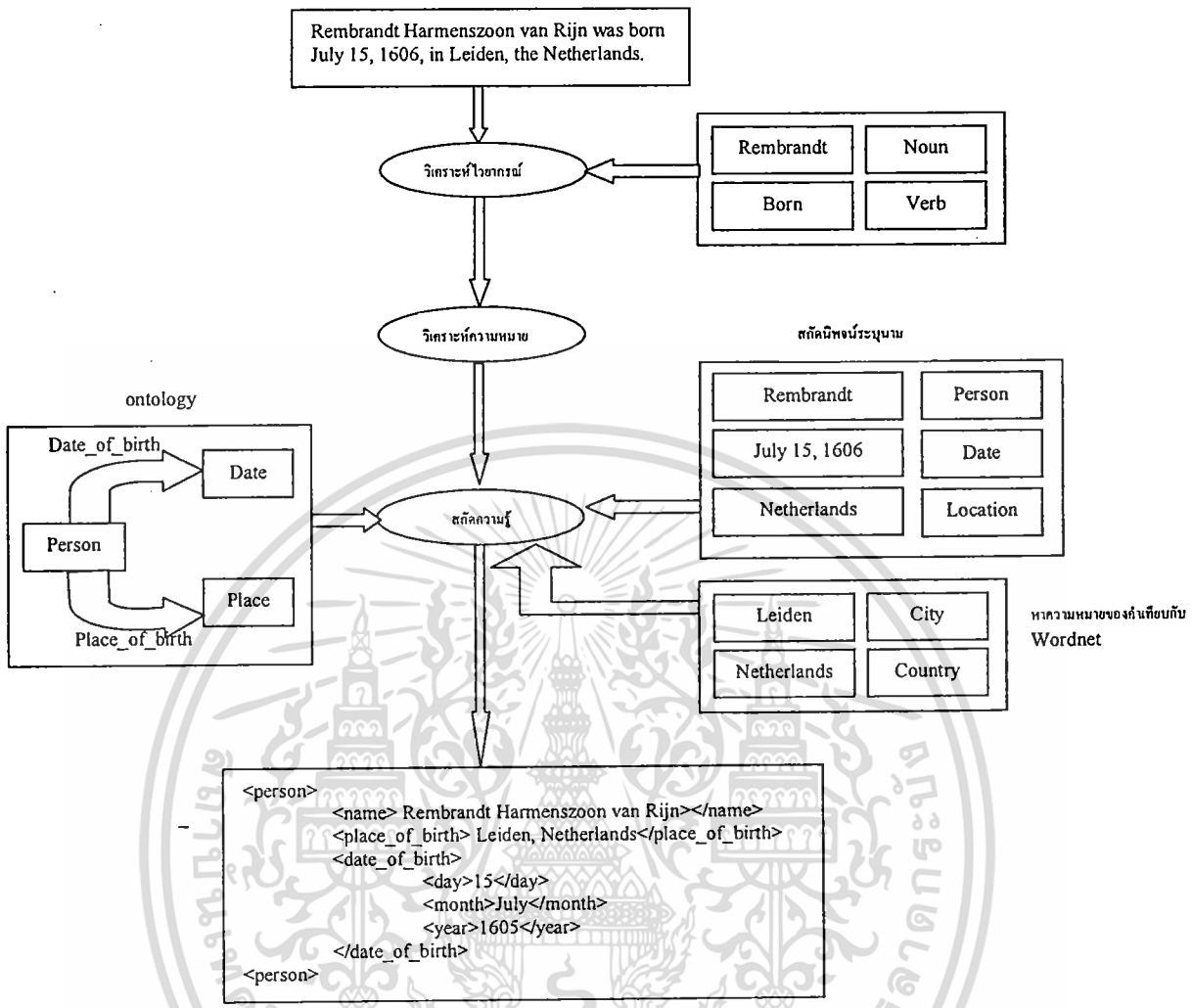
ในงานวิจัยนี้ จะไม่มีการกำหนดโดเมนให้กับข้อมูลที่ต้องการกำหนด นั่นคือสามารถใช้ได้กับเอกสารทั่ว ๆ ไป แต่ถ้าต้องการให้กำกับความหมายในเรื่องที่เฉพาะนั้นจะต้องทำการปรับเปลี่ยนออนโทโลจีให้เหมาะสมกับโดเมนที่ต้องการกำกับความหมาย

Alani, et al. (2003) ได้นำเสนอเครื่องมือในการสกัดความรู้จากเว็บเอกสารที่เรียกว่า Artequake โดยผู้ใช้ใส่ชื่อศิลปินที่ต้องการสืบค้น จากนั้นระบบจะทำการประมวลผลเพื่อแสดงผลลัพธ์ข้อมูลรายละเอียดเกี่ยวกับศิลปินผู้นั้นในลักษณะของการสรุปรายละเอียดที่เกี่ยวกับศิลปินท่านนั้นให้กับผู้ใช้ ทำให้ผู้ใช้งานสามารถได้ข้อมูลศิลปินท่านนั้นอย่างละเอียดโดยที่ไม่ต้องเข้าไปค้นหาหลายเว็บไซต์ ในส่วนของการประมวลผลนั้น จะทำการสกัดความรู้จากเอกสารที่เกี่ยวข้องที่ได้ถูกเลือกโดยพิจารณาเปรียบเทียบกับข้อมูลศิลปินในเว็บไซต์ www.ibiblio.org/wm/paint โดยในการสกัดความรู้นั้นได้มีการนำออนโทโลจีที่แสดงคอนเซ็ปและความสัมพันธ์ของโดเมนที่ต้องการสกัดมาช่วยทำให้การพิจารณาเลือกข้อมูลที่สกัดมีความถูกต้องมากยิ่งขึ้น แล้วนำข้อมูลที่สกัดได้รวบรวมแสดงให้กับผู้ใช้งานในรูปแบบของการสรุปรายละเอียดของศิลปินที่ผู้ใช้ได้ทำการเลือกไว้

ข้างต้น นอกจากนั้นแล้วข้อมูลที่สกัดได้ยังสามารถถูกนำกลับไปเป็นอินสแตนซ์ของออนโทโลจีอีกด้วย เพื่อที่เมื่อมีการค้นหาข้อมูลศิลปินที่เคยสืบค้นแล้ว ระบบเพียงแต่นำข้อมูลที่ได้ทำการสกัดเรียบร้อยแล้วที่จัดเก็บในออนโทโลจีมาแสดงให้กับผู้ใช้งานได้เลย

สำหรับออนโทโลจีในงานวิจัยนี้ เป็นรูปแบบการแทนความหมายของโดเมนที่เราสนใจ ซึ่งเครื่องสามารถเข้าใจได้ โดยออนโทโลจีที่สร้างขึ้นนี้เป็นออนโทโลจีที่แทนคอนเซ็ป รายละอียดคุณลักษณะเกี่ยวกับข้อมูลศิลปินและความสัมพันธ์ระหว่างคอนเซ็ป โดยคอนเซ็ปและความสัมพันธ์จะประกอบด้วย ผลงาน สิ่งประดิษฐ์ สถานที่ ครอบครัว ข้อมูลส่วนบุคคล ความสัมพันธ์กับศิลปินบุคคลอื่น เป็นต้น

ภาพที่ 2.6 แสดงขั้นตอนของการสกัดความรู้จากเอกสาร โดยเมื่อระบบได้รวบรวมเอกสารที่เกี่ยวข้องกับแล้ว จะพิจารณาข้อความในเอกสารทีละย่อหน้า โดยนำข้อความนั้นไปวิเคราะห์ไวยากรณ์ และความหมาย ซึ่งในการวิเคราะห์ไวยากรณ์จะพิจารณาว่าคำใดทำหน้าที่เป็นประธาน กริยา กรรม และส่วนประกอบอื่น ๆ ของประโยค และสกัดนิพจน์ระบุนามจากกลุ่มคำในประโยคที่ทำหน้าที่เป็นชื่อคน ชื่อสถานที่ วันเกิด เป็นต้น นอกจากนั้นยังนำคำเหล่านั้นไปหาความหมายเทียบกับ WordNet โดยในการสกัดความรู้นั้นจะอาศัยออนโทโลจีในการบอกว่าข้อมูลที่ต้องการสกัดประกอบด้วยอะไรบ้างโดยแสดงในรูปของคอนเซ็ป อีกทั้งในออนโทโลจียังแสดงความสัมพันธ์ระหว่างคอนเซ็ป เช่น คอนเซ็ปบุคคล (Person) สัมพันธ์กับคอนเซ็ปวันที่ ผ่านความสัมพันธ์ มีวันเกิด เมื่อสกัดความรู้จากข้อความแล้วจะจัดเก็บในรูปแบบของ XML และ จัดเก็บในรูปแบบของออนโทโลจี ในขั้นตอนสุดท้ายจะนำข้อความที่ผ่านการวิเคราะห์แล้วมาแสดงในรูปแบบของการสรุปตามเทมเพลตของการแสดงผลลัพธ์ของข้อมูลที่ได้เตรียมไว้แล้ว



ภาพที่ 2.6 แสดงขั้นตอนการสกัดความรู้จากเอกสารในงานวิจัย Alani, et al. (2003)

จากการวิจัยนี้ พบว่ามีการใช้ออนโทโลจีเป็นที่ปรึกษาในการสกัดความรู้จากข้อความ โดยระบุว่ามีข้อมูลอะไรที่ต้องการสกัดและข้อมูลเหล่านั้นมีความสัมพันธ์กันอย่างไร ซึ่งจะคล้ายกับการทำงานของการสกัดข้อมูลแบบเดิมที่ต้องอาศัยเฟรมสล็อตในการระบุข้อมูลที่ต้องการสกัด แต่ในงานวิจัยนี้ใช้ความสามารถของออนโทโลจีในการระบุข้อมูลแทน

Laclavik, et al. (2009) พัฒนาเครื่องมือสำหรับการกำกับความหมายให้กับเว็บเอกสารแบบกึ่งอัตโนมัติที่เรียกว่า OnTeA โดยการใช้แพทเทิร์นในการระบุกลุ่มคำหรือคำที่ต้องการกำกับความหมายให้ เอกสารที่ต้องการกำกับจะถูกระบุข้อมูลที่ต้องการกำกับโดยเทียบกับแพทเทิร์นที่กำหนดไว้ใน Pattern Ontology ถ้าข้อมูลนั้นปรากฏเป็นอินสแตนซ์ในออนโทโลจี ก็จะทำระบุข้อมูลนั้นด้วยคอนเซ็ปของอินสแตนซ์นั้น กรณีที่ไม่พบในอินสแตนซ์ก็จะใช้แพทเทิร์นเป็นตัวระบุ

ว่าคอนเซ็ปนั้นควรมีอินสแตนซ์ที่มีรูปแบบของคำเป็นลักษณะใด เช่น ในการระบุชื่อสถานที่ สามารถกำหนดแพทเทิร์นได้ว่า (in|near) +([A-Z][a-z]+) เมื่อระบบเจอคำว่า in หรือ near แล้วตามด้วยตัวอักษรขึ้นต้นตัวใหญ่ ก็จะระบุได้ว่า กลุ่มนั้นคืออินสแตนซ์ของคอนเซ็ปสถานที่ ซึ่งในการกำหนดแพทเทิร์นนั้นจะอาศัยกฎของ regular expression

ตัวอย่างเช่น

Rembrandt Harmenszoon Van Rijn was born July 15, 1606, in Leiden, the Netherlands.

แพทเทิร์น :

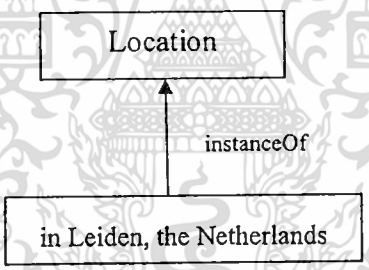
Location : (in|near) +([A-Z][a-z]+[,]*[A-Z][a-z]*) -> in Leiden, the Netherlands

Date : [A-z][a-z][0-9]+[,][0-9]+ -> July 15, 1606

Output : Location – in Leiden, the Netherlands

Date - July 15, 1606

และจากผลลัพธ์ที่ได้จะถูกนำเข้าไปเพิ่มเป็นอินสแตนซ์ของออนโทโลจี แสดงดังภาพที่ 2.7



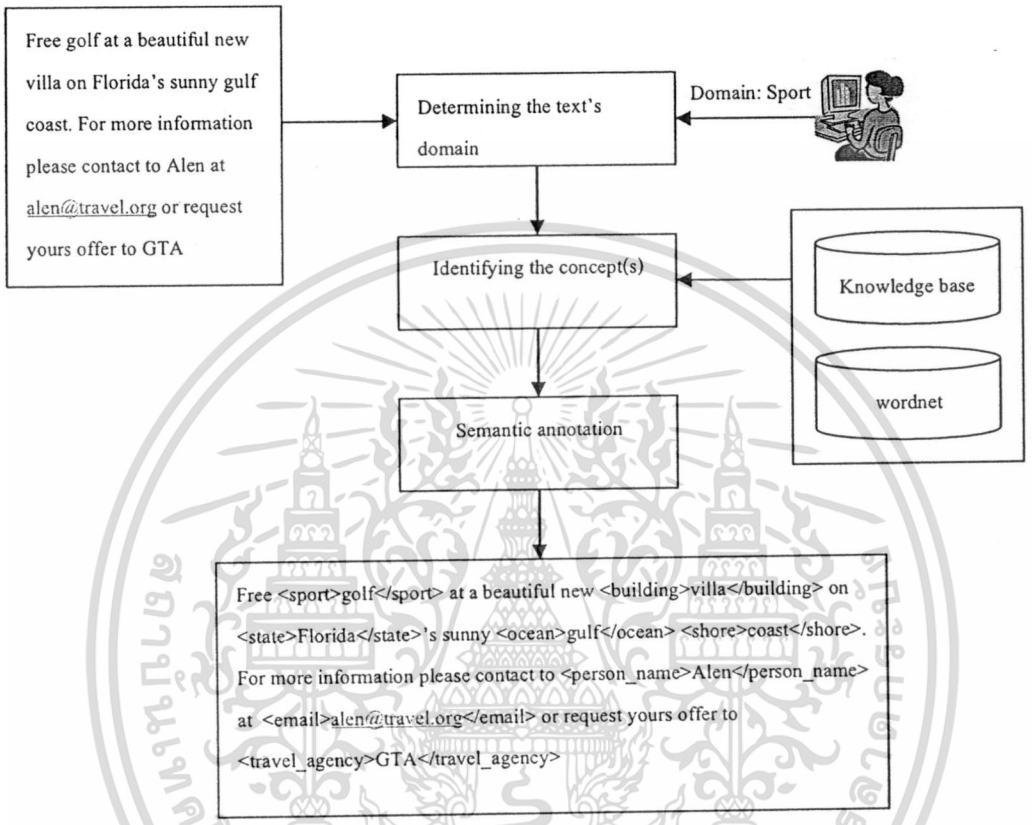
ภาพที่ 2.7 แสดงการเพิ่มอินสแตนซ์ในออนโทโลจี

จากงานวิจัยนี้พบว่าในการกำกับความหมายให้กับข้อมูลนั้นจะขึ้นอยู่กับแพทเทิร์นที่กำหนดว่าแต่ละคอนเซ็ปควรมีอินสแตนซ์เป็นข้อมูลลักษณะเช่นไร ถ้ามีการกำหนดแพทเทิร์นที่ดีก็จะทำให้ได้ข้อมูลที่มีการกำกับคอนเซ็ปที่ถูกต้อง

Yasrebi and Mohsenadeh (2009) เสนอวิธีการกำกับความหมายแบบกึ่งอัตโนมัติโดยอาศัยฐานความรู้ที่จัดเก็บในรูปแบบออนโทโลจีในการพิจารณาความหมายของคำที่ทำการกำกับ นอกจากนั้นยังใช้ wordnet ในการพิจารณาร่วมด้วยในกรณีที่คำที่พิจารณาไม่ปรากฏในออนโทโลจี โดยได้ทำการพัฒนาเครื่องมือที่ใช้ในการกำกับความหมาย

จากภาพที่ 2.8 เอกสารจะถูกกำหนดว่าเป็นเอกสารประเภทใด เช่น “travel” “location” เป็นต้น โดยผู้ใช้งานระบบเป็นผู้กำหนด แล้วทำการระบุคอนเซ็ปให้กับข้อมูลในเอกสารที่เกี่ยวข้อง

กับกลุ่มของเอกสารที่ถูกระบุข้างต้น โดยทำการเปรียบเทียบจากฐานความรู้ ในกรณีที่ไม่พบในฐานความรู้ ข้อมูลนั้นจะถูกนำไปวิเคราะห์กับ wordnet ว่าคำนั้นมีคอนเซ็ปที่เกี่ยวข้องกับกลุ่มของเอกสารหรือไม่



ภาพที่ 2.8 แสดงสถาปัตยกรรมระบบของ Yasrebi and Mohsenadeh (2009)

ในงานวิจัยนี้ได้มีการใช้ WordNet เพื่อนำมาวิเคราะห์คำที่ไม่รู้จัก เพื่อให้ระบบสามารถวิเคราะห์คำเหล่านั้นได้ แต่อย่างไรก็ตามคำใหม่นั้นจะต้องมีความหมายอยู่ในกลุ่มคำที่ปรากฏในออนโทโลจีด้วย

2.2.4 เทคนิคการวิเคราะห์ข้อความโดยการฐานความรู้อื่น

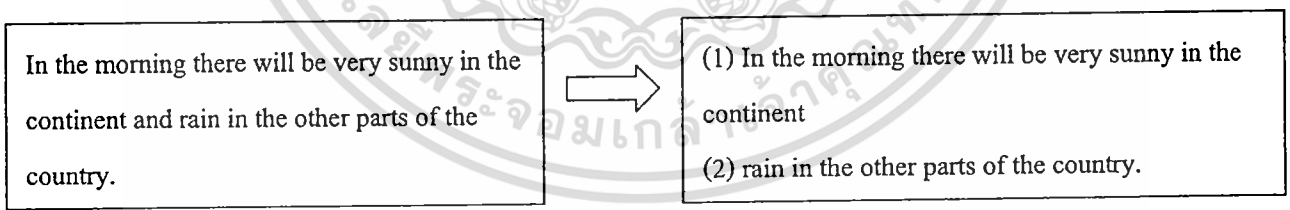
Mestrovic, et al. (2007) ได้ใช้เทคนิคของการวิเคราะห์ความหมายในการสกัดข้อมูลจากเอกสารแล้วทำการจัดเก็บข้อมูลที่สกัดได้นั้นไว้ในรูปแบบของเฟรม ในการวิเคราะห์ความหมายนั้นจะอาศัย frame logic หรือ F-logic ในการแทนข้อมูลในพจนานุกรม กลุ่มของข้อมูล และผลลัพธ์ที่ปรากฏในเฟรม ซึ่งรูปแบบจะอยู่ในลักษณะ object[attribute->value] และข้อมูลที่ใช้ในงานวิจัยนี้เป็นเอกสารที่เกี่ยวกับการพยากรณ์อากาศ

การสกัดข้อมูลในงานวิจัยนี้มีขั้นตอนดังนี้

1. จัดเอกสารว่าเป็นเอกสารกลุ่มใด (weather forecast, sea weather forecast, bio weather forecast, meteorology, the state of the river, wind, temperature, place, time, description, irrelevant) และแต่ละกลุ่มก็จะประกอบด้วยกลุ่มย่อย ๆ ซึ่งแต่ละกลุ่มย่อยจะอธิบายรายละเอียดส่วนประกอบของแต่ละกลุ่ม เช่น กลุ่ม "wind" ประกอบด้วยกลุ่มย่อยคือ "direction" "intensity" "name" "blowing" โดยอาจกล่าวได้ว่ากลุ่มย่อยคือส่วนทำหน้าที่สล็อตไว้อธิบายรายละเอียดของแต่ละกลุ่มและเป็นผลลัพธ์ของการแสดงข้อมูลที่สกัดได้ โดยแต่ละกลุ่มนั้นสามารถมีความสัมพันธ์ระหว่างกลุ่มได้ในการจัดกลุ่มให้กับเอกสารนั้นพิจารณาจากกลุ่มคำที่ปรากฏในเอกสารเปรียบเทียบกับพจนานุกรมที่สร้างขึ้นในแต่ละกลุ่มข้างต้น จากภาพที่ 2.9 แสดงกลุ่มของข้อความที่ถูกจัดกลุ่มให้อยู่ภายใต้กลุ่ม "weather forecast"

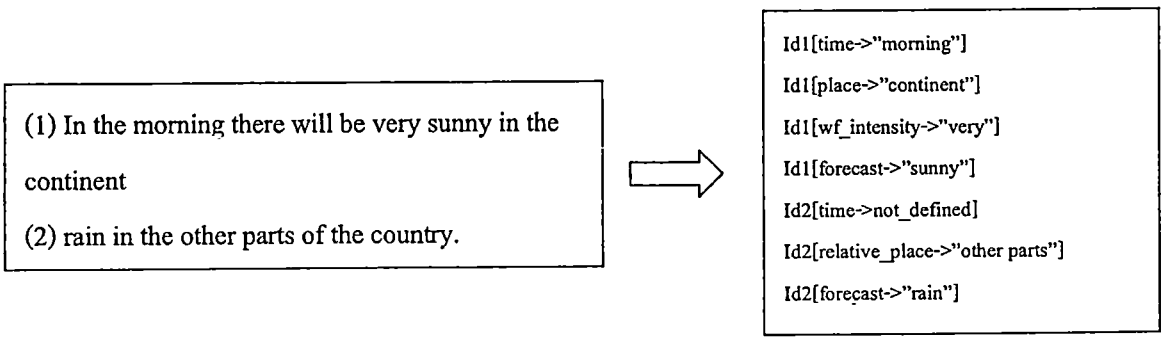
ภาพที่ 2.9 แสดงกลุ่มของข้อความที่ถูกจัดกลุ่มให้อยู่ภายใต้กลุ่ม "weather forecast"

2. วิเคราะห์ความหมายและสกัดข้อมูลจากเอกสารแล้วนำไปใส่ในสล็อตของเฟรม ในการวิเคราะห์ความหมายนั้นจะทำการวิเคราะห์ทีละประโยค และในแต่ละประโยคนั้นจะถูกแบ่งออกเป็นประโยคย่อย โดยพิจารณาจากการใช้เครื่องหมายคอมมา และคำเชื่อม จากตัวอย่างข้างต้นสามารถแบ่งออกได้เป็น 2 กลุ่มประโยคย่อย แสดงดังภาพที่ 2.10



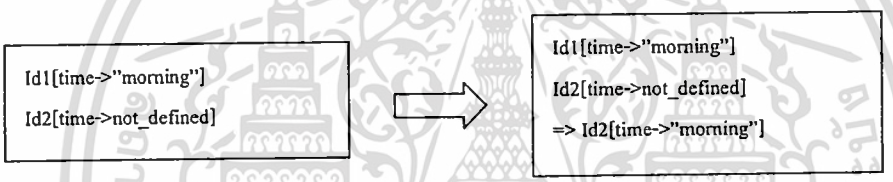
ภาพที่ 2.10 แสดงการแบ่งประโยคออกเป็นประโยคย่อย

เมื่อแบ่งประโยคออกเป็นประโยคย่อยแล้ว จากนั้นระบุบทบาทของแต่ละคำที่เกี่ยวข้องในข้อความตามพจนานุกรมที่สร้างขึ้น แสดงดังภาพที่ 2.11



ภาพที่ 2.11 ระบุบทบาทของคำ

3. ปรับปรุงข้อมูลที่สกัดได้ให้มีความสมบูรณ์ เนื่องจากว่าข้อมูลบางอย่างที่สกัดได้มีการละความหมายไว้ แต่ยังสามารถอ้างอิงได้จากข้อมูลในประโยคเดียวกัน จากภาพที่ 2.12 พบว่าข้อมูลเวลาไม่ปรากฏในประโยคย่อยส่วนที่ 2 แต่ปรากฏในส่วนที่ 1 ดังนั้นในประโยคย่อยที่ 2 สามารถอ้างอิงไปใช้ข้อมูลในประโยคย่อยที่ 1 ได้



ภาพที่ 2.12 ปรับปรุงข้อมูลให้สมบูรณ์

ในงานวิจัยนี้ได้มีการใช้เทคนิคของการวิเคราะห์ความหมายของคำในประโยคเพื่อทำให้การสกัดข้อมูลมีประสิทธิภาพมากขึ้น โดยอาศัยพจนานุกรมในการวิเคราะห์ความหมาย ซึ่งพบว่าการที่จะทำให้พจนานุกรมนั้นประกอบด้วยคำที่ครอบคลุมทั้งหมดนั้นเป็นเรื่องที่ยุ่งยาก ทำให้คำที่ไม่ปรากฏในพจนานุกรมจะไม่สามารถระบุความหมายได้ ถึงแม้ว่าคำนั้นจะเป็นคำที่เกี่ยวข้องกับข้อมูลที่ต้องการสกัดก็ตาม

Zeni, et al (2007) ได้พัฒนาระบบสำหรับการกำกับความหมายให้กับเอกสารโฆษณาที่พักอาศัยโดยจะกำกับข้อความในส่วนของสถานที่ ราคา สถานที่ติดต่อ สิ่งอำนวยความสะดวก ประเภทของที่พัก โดยจัดเก็บข้อมูลที่ทำการกำกับไว้ในฐานข้อมูล เพื่อสามารถนำข้อมูลเหล่านั้นไปใช้ในงานอื่นต่อในอนาคต การกำกับเอกสารนี้ประกอบด้วย 3 ขั้นตอน ได้แก่

1. Parse จะทำการวิเคราะห์ข้อมูลในเอกสารให้อยู่ในรูปแบบต้นไม้ ที่จะทำการแบ่งส่วนของคำหรือกลุ่มคำในเอกสาร ด้วยกฎ โดยกฎนั้นจะทำการจำแนกข้อมูลในกลุ่มของอีเมล์ ชื่อเว็บไซต์ จำนวนเงิน วันที่ เวลา เป็นต้น ในการแจกแจงข้อมูลให้อยู่ในรูปแบบของต้นไม้นี้ได้ใช้ TXL parsing engine เป็นเครื่องมือในการทำงาน

2. Markup จะทำการกำกับความหมายของเอกสารตามสกีมาที่กำหนดไว้ว่าต้องการกำกับข้อมูลอะไรบ้าง โดยแต่ละข้อมูลนั้นจะมีตัวระบุหรือคำสำคัญเพื่อใช้กำหนดสิ่งที่ต้องการกำกับ

3. Mapping เมื่อกำกับข้อมูลจากข้อ 2 แล้ว ระบบจะทำการแม็ปข้อมูลนั้นเข้ากับเทมเพลต แล้วจัดเก็บในฐานข้อมูล

ในงานวิจัยนี้ใช้คำสำคัญเป็นตัวระบุข้อมูลที่ต้องการจัดเก็บ ซึ่งสามารถปรับเปลี่ยนใช้กับงานในโดเมนอื่นได้ง่าย

งานวิจัยดังที่ได้กล่าวมาข้างต้นเป็นงานวิจัยที่เกี่ยวกับการวิเคราะห์เนื้อความเอกสาร ซึ่งมีความแตกต่างกันในด้านเทคนิคและเทคโนโลยีที่เลือกใช้ในการวิเคราะห์เนื้อความ โดยในบทที่ 3 จะทำการเปรียบเทียบแต่ละงานวิจัยพร้อมเสนอแนวทางปรับปรุงในแต่ละงานวิจัย

บทที่ 3
เปรียบเทียบเทคนิคที่เกี่ยวข้อง

จากการศึกษาเทคนิคการวิเคราะห์ข้อความเว็บเอกสาร ซึ่งออกเป็น 4 กลุ่ม ได้แก่ การวิเคราะห์ข้อความด้วยวิธีการเปรียบเทียบรูปแบบ (Pattern-Matching) การวิเคราะห์ข้อความด้วยวิธีการเรียนรู้ของเครื่องจักร (Machine Learning) การวิเคราะห์ข้อความโดยการใช้ออนโทโลจี และการวิเคราะห์ข้อความโดยการใช้ฐานความรู้อื่น ดังนั้นในบทนี้จึงได้นำเสนอการเปรียบเทียบเทคนิคเหล่านี้ในแต่ละงานวิจัย

ตารางที่ 3.1 แสดงรายละเอียดอธิบายเทคนิคแต่ละวิธีแบบย่อ ๆ อินพุตของระบบ ผลลัพธ์ของระบบ และโดเมนที่ใช้ในการทำงานของแต่ละเทคนิค งานวิจัย (Riloff and Schmelzenbach, 1998) เป็นการวิเคราะห์ข้อความด้วยวิธีการเปรียบเทียบรูปแบบ งานวิจัย (Tellez-Valero et al., 2009) เป็นการวิเคราะห์ข้อความด้วยวิธีการเรียนรู้ของเครื่องจักร งานวิจัย (Alani et al., 2002, Popov et al., 2003, and Laclavik et al., 2009) เป็นการวิเคราะห์ข้อความโดยการใช้ออนโทโลจี และ งานวิจัย (Mestrovic et al., 2007, Kiyavitskaya et al., 2009) เป็นการวิเคราะห์ข้อความโดยการใช้ฐานความรู้อื่น

พบว่างานวิจัยทางด้านการวิเคราะห์ข้อความเอกสารโดยส่วนใหญ่แล้วจะทำงานกับโดเมนเฉพาะ โดยจะเป็นโดเมนที่มีขอบเขตจำกัดที่สามารถระบุถึงสิ่งที่ต้องการวิเคราะห์ๆ ได้อย่างแน่ชัด เช่นโดเมนข่าวก่อการร้าย ข่าวสภาพอากาศ ภัยธรรมชาติ โฆษณาที่พักอาศัย ประวัติศิลปิน เป็นต้น แต่อย่างไรก็ตามได้มีความพยายามจะทำให้สามารถวิเคราะห์ได้กับโดเมนทั่ว ๆ ไป ซึ่งก็ต้องอาศัยฐานความรู้ขนาดใหญ่ที่จะทำให้คอมพิวเตอร์นั้นสามารถวิเคราะห์ข้อความได้ทุกโดเมน ซึ่งในงานวิจัย (Alani et al., 2002 and Laclavik et al., 2009) นั้นสามารถขยายให้ครอบคลุมกับเนื้อหาในโดเมนอื่นได้ โดยการขยายฐานความรู้ให้มากยิ่งขึ้น เนื่องจากในงานวิจัยเหล่านี้มีความยืดหยุ่น มีการแยกส่วนของฐานความรู้แยกจากระบบ

ในปัจจุบันฐานความรู้ที่นิยมใช้ในงานวิจัยด้านนี้คือ ออนโทโลจี โดยความรู้ของแต่ละโดเมนจะแทนด้วยออนโทโลจี ที่กำหนดขอบเขตของเนื้อหาข้อมูลของโดเมนนั้น งานวิจัย (Popov et al., 2003, Laclavik et al., 2009, Yasrebi and Mohsenadeh, 2009) ใช้ออนโทโลจีเพื่อกำหนดขอบเขตของเนื้อหาของโดเมน และระบุข้อมูลที่ต้องการวิเคราะห์ ในขณะที่งานวิจัย (Mestrovic et al., 2007, Kiyavitskaya et al., 2009) ใช้รูปแบบของฐานความรู้ในลักษณะอื่นในการกำหนดขอบเขตของข้อมูลโดเมน โดยใช้กฎในการระบุข้อมูล ซึ่งจะทำให้นำไปใช้งานกับระบบอื่นค่อนข้างยากในขณะที่ออนโทโลจีเป็นฐานความรู้ที่สามารถนำไปใช้กับงานอื่น ๆ ได้อย่างสะดวก

งานวิจัย (Alani et al., 2002, Popov et al., 2003, and Tellez-Valero et al., 2009) ได้มีการใช้เทคนิคการประมวลผลภาษาธรรมชาติเข้ามาช่วยในการวิเคราะห์ข้อความเอกสาร ซึ่งทำให้การวิเคราะห์นี้ได้คำนึงถึงโครงสร้างประโยคและความหมายของข้อความที่วิเคราะห์ และเป็นส่วนที่สำคัญในการวิเคราะห์ข้อความเอกสาร

ตารางที่ 3.2 แสดงการเปรียบเทียบข้อเด่นและข้อด้อยของงานวิจัยที่ศึกษา พบว่าเทคนิคที่ใช้ในงานวิจัยต่าง ๆ ยังต้องอาศัยผู้เชี่ยวชาญในการตรวจสอบความถูกต้อง การสร้างฐานความรู้ การสร้างกฎต่าง ๆ สำหรับระบบ ซึ่งเป็นการยากที่จะทำให้คอมพิวเตอร์สามารถทำงานได้โดยทั้งหมด แต่อย่างไรก็ตามงานวิจัยเหล่านั้นก็พยายามที่จะทำให้ระบบสามารถทำงานได้ด้วยตนเองได้มากที่สุดและอาศัยมนุษย์ให้น้อยที่สุด

ข้อเด่นของงานวิจัย (Alani et al., 2002, Yasrebi and Mohsenadeh, 2009) ที่แตกต่างจากงานอื่นคือได้มีการใช้ WordNet เข้ามาช่วยในการวิเคราะห์ความหมายของคำที่เกี่ยวข้องสัมพันธ์กันในลักษณะต่าง ๆ ทำให้เมื่อปรากฏคำที่ไม่รู้จัก ระบบจะสามารถวิเคราะห์ได้ว่าเกี่ยวข้องกับข้อมูลในฐานความรู้ที่ได้กำหนดไว้ล่วงหน้าหรือไม่ แต่ก็ทำให้ระบบต้องทำงานเพิ่มขึ้นและใช้เวลาในการประมวลผลมากขึ้นด้วยเช่นกัน

Kiyavitskaya et al. (2009) ได้พยายามทำให้การใช้ทรัพยากรของระบบน้อยลง โดยการใช้ TXL engine แทนการใช้ออนโทโลจี ทำให้สามารถทำงานได้เร็วกว่า แต่อย่างไรก็ตามก็ยังต้องอาศัยกฎในการวิเคราะห์ข้อความ

Popov et al. (2003) ได้สร้างระบบที่สามารถใช้งานได้กับทุกโดเมน แต่ก็ยังเป็นการวิเคราะห์ในส่วนกว้าง ๆ ไม่เฉพาะเจาะจง ในขณะที่งานวิจัยอื่นจะทำงานกับโดเมนเฉพาะ ที่ให้ผลลัพธ์เฉพาะเจาะจงในแต่ละโดเมน

ตารางที่ 3.1 แสดงรายละเอียดอธิบายเทคนิคแต่ละวิธีแบบย่อ ๆ อินพุตของระบบ ผลลัพธ์ของระบบ และโดเมนที่ใช้ในการทำงานของแต่ละเทคนิค

งานวิจัย	วิธีการ	อินพุต	ผลลัพธ์	โดเมน
An Empirical Approach to Concept Case Frame Acquisition (Riloff and Schmelzenbach, 1998)	สร้างเฟรมอัตโนมัติในการสกัดข้อมูล โดยแต่ละเฟรมนั้นจะกำหนดสล็อตและรูปแบบของข้อมูลที่ต้องการสกัดในแต่ละสล็อต ซึ่งในการสร้างเฟรมนั้นจะอ้างอิงมาจากแพทเทิร์นที่เรียนรู้มาจากเอกสารตัวอย่าง ส่วนขั้นตอนของการสกัดข้อมูลนั้นจะนำประโยคเทียบกับเฟรมต่างๆ เพื่อเลือกว่าประโยคนั้นเหมาะสมกับแพทเทิร์นใด	เอกสารข่าว, ข้อมูล Domain role, Semantic Category	เฟรมสล็อตของข้อมูลที่ต้องการสกัดจากเอกสาร	ข่าวก่อการร้าย
Using Machine Learning for Extracting Information from Natural Disaster News Reports (Tellez-Valero et al., 2009)	สกัดข้อมูลจากเอกสารโดเมน โดยเลือกว่าเอกสารนั้นอยู่ในโดเมนที่ต้องการสกัดหรือไม่ จากนั้นจึงทำการวิเคราะห์ข้อมูลที่ต้องการสกัด โดยการใช้ regular expression ในการสกัดข้อมูลให้สอดคล้องกับเฟรมที่ระบุ	เอกสาร, รูปแบบเฟรมที่ต้องการสกัด	เฟรมสล็อตที่ประกอบด้วยข้อมูลที่สกัดจากเอกสาร	ข่าวภัยธรรมชาติ
Toward Semantic Web Information Extraction (Popov et al., 2003)	กำกับความหมายให้กับข้อความในเอกสารโดย แบ่งข้อความในเอกสารออกเป็นโทเคน วิเคราะห์หน้าที่ของคำ สกัดข้อมูลที่ต้องการจากเอกสาร แล้วกำกับความหมายของคำตามออนโทโลจีที่กำหนด	ออนโทโลจี เอกสารที่ต้องการกำกับความหมาย	เอกสารที่ถูกกำกับความหมาย	ทุกโดเมน

ตารางที่ 3.1 แสดงรายละเอียดอธิบายเทคนิคแต่ละวิธีแบบย่อ ๆ อินพุตของระบบ ผลลัพธ์ของระบบ และโดเมนที่ใช้ในการทำงานของแต่ละเทคนิค (ต่อ)

งานวิจัย	วิธีการ	อินพุต	ผลลัพธ์	โดเมน
Automatic Ontology-Based Knowledge Extraction from Web Documents (Alani et al., 2002)	รวบรวมเอกสารจากเว็บไซต์ตามชื่อศิลปิน แล้วทำการเลือกเอกสารที่เกี่ยวข้องโดยวัดค่าความคล้ายคลึงเทียบกับเอกสารที่กำหนด จากนั้นนำเอกสารที่เลือกมาวิเคราะห์ทีละประโยค โดยทำการวิเคราะห์หน้าที่ของคำและหากลุ่มคำที่ทำหน้าที่เป็นประธาน กริยาและกรรม แล้วหาความสัมพันธ์ที่เกิดขึ้นภายในประโยคตามความสัมพันธ์ที่กำหนดในออนโทโลจี โดยใช้ GATE และ WordNet ในการพิจารณาความหมายของคำให้สอดคล้องกับความสัมพันธ์ แล้วสร้างข้อมูลที่เกี่ยวข้องกับความสัมพันธ์เป็นอินสแตนซ์ในออนโทโลจี	เว็บเอกสารข้อมูลศิลปิน, โดเมนออนโทโลจี, wordnet	อินสแตนซ์ในโดเมนออนโทโลจี โดยอินสแตนซ์ได้จากข้อความที่ปรากฏในเอกสาร, เอกสารชีวประวัติของศิลปินที่สร้างใหม่โดยรวบรวมจากข้อมูลหลาย ๆ เว็บไซต์	ข้อมูลประวัติศิลปิน
ONTEA : Platform for pattern based automated semantic annotation (Laclavik et al., 2009)	กำกับความหมายให้กับข้อความในเอกสารโดยการใช้ regular expression patterns ที่กำหนดไว้ใน pattern ontology ในการเลือกข้อความให้สอดคล้องกับความหมายที่กำหนดใน domain ontology	โดเมนออนโทโลจี, แพทเทิร์นออนโทโลจี, เว็บเอกสาร	อินสแตนซ์ของเว็บเอกสารประกอบด้วยคุณสมบัติรายละเอียดภายในเอกสาร และเพิ่มอินสแตนซ์นั้นในโดเมนออนโทโลจี	การสมัครงาน

ตารางที่ 3.1 แสดงรายละเอียดอธิบายเทคนิคแต่ละวิธีแบบย่อ ๆ อินพุตของระบบ ผลลัพธ์ของระบบ และโดเมนที่ใช้ในการทำงานของแต่ละเทคนิค (ต่อ)

งานวิจัย	วิธีการ	อินพุต	ผลลัพธ์	โดเมน
Semi-Automatic Approach for Semantic Annotation (Yasrebi and Mohsenadeh, 2009)	ใช้ฐานความรู้ที่ระบุข้อมูลภายในโดเมน และ Wordnet ในการกำกับความหมายให้กับคำในเอกสาร โดยนำคำในเอกสารมาพิจารณากับข้อมูลในฐานความรู้ ถ้าไม่มีก็จะใช้ wordnet ในการวิเคราะห์ว่าคำนั้นสัมพันธ์กับคำใดในฐานความรู้หรือไม่ แล้วทำการกำกับความหมายตามคอนเซ็ปของคำนั้น	เอกสาร wordnet ฐานความรู้	เอกสารที่กำกับความหมาย	กีฬา
Weather Forecast Data Semantic Analysis in F-logic (Mestrovic et al., 2007)	กำหนดกลุ่มข้อมูลพยากรณ์ให้กับข้อความที่เป็นอินพุต วิเคราะห์ข้อความทีละประโยค โดยเปรียบเทียบข้อความกับดิกชันนารีที่สร้างขึ้นเองเพื่อกำหนดคอนเซ็ปของข้อความในประโยค แล้วจัดเก็บข้อมูลลงในสล็อตตามรูปแบบเฟรมของแต่ละกลุ่มข้อมูลพยากรณ์	Semantic categories, Semantic dictionary, Output template, เว็บเอกสาร	Frame-slot ในรูปแบบของ F-logic	พยากรณ์อากาศ
Cerno : Light-weight tool support for semantic annotation of textual documents (Kiyavitskaya et al., 2009)	วิเคราะห์โครงสร้างเอกสารตาม context free grammar ด้วย TXL engine จากนั้นกำกับความหมายให้กับกลุ่มคำตาม annotation schema แล้วจัดเก็บข้อมูลลงฐานข้อมูลตาม database schema	Context free grammar, Annotation schema, Database schema Semantic model	จัดเก็บข้อมูลที่กำกับความหมายลงในฐานข้อมูล	โฆษณาที่พัก

ตารางที่ 3.2 ตารางแสดงการเปรียบเทียบงานวิจัยที่ศึกษา

งานวิจัย	ข้อเด่น	ข้อด้อย	ข้อเสนอแนะในการปรับปรุง
An Empirical Approach to Concept Case Frame Acquisition (Riloff and Schmelzenbach, 1998)	- สร้างเฟรมสำหรับการสกัดข้อมูลโดยอัตโนมัติ - เหมาะกับโดเมนที่ไม่มีความหลากหลายของประโยคมากนัก	- ต้องอาศัยผู้เชี่ยวชาญในเลือกแพทเทิร์นที่เหมาะสมสำหรับโดเมน - แม้ว่าจะเป็นการสร้างเฟรมโดยอัตโนมัติ แต่อย่างไรก็ตาม ก็ต้องมีการกำหนดชื่อสล็อตไว้ล่วงหน้า - ถ้ารูปแบบประโยคมีความหลากหลายและซับซ้อนการสร้างแพทเทิร์นก็จะยุ่งยากไปด้วย	ในการสกัดข้อมูลนั้นจะขึ้นอยู่กับแพทเทิร์นที่ได้เรียนรู้จากข้อมูลตัวอย่าง และเป็นแพทเทิร์นนั้นเป็นรูปแบบสำหรับสกัดข้อมูลที่ไม่ซับซ้อน ดังนั้นควรปรับปรุงการกำหนดแพทเทิร์นให้ครอบคลุมกับรูปแบบประโยคที่หลากหลาย
Using Machine Learning for Extracting Information from Natural Disaster News Reports (Tellez-Valero et al., 2009)	- ระบบสามารถเลือกเอกสารที่เกี่ยวข้องกับโดเมนได้โดยอัตโนมัติ - ในขั้นตอนการสกัดข้อมูลนั้นสามารถทำงานได้เร็ว เนื่องจากการวิเคราะห์ข้อมูลที่ต้องการสกัดนั้นใช้รูปแบบตาม regular expression	- ข้อมูลที่สกัดนั้นเป็นข้อมูลที่ไม่ซับซ้อน - ไม่มีการวิเคราะห์โครงสร้างของประโยคในการสกัดข้อมูล	เนื่องจากในงานวิจัยนี้ทำงานกับโดเมนภัยธรรมชาติ และในการสกัดข้อมูลที่สอดคล้องกับโดเมนนั้น ใช้เพียงรูปแบบที่กำหนดตาม regular expression ซึ่งทำให้รายละเอียดของข้อมูลที่สกัดได้นั้นมีขอบเขตที่ค่อนข้างจำกัด ดังนั้นควรจะปรับปรุงโดยนำฐานความรู้เข้ามาช่วยในการวิเคราะห์ข้อมูลที่ต้องการสกัดเพิ่มเติม

ตารางที่ 3.2 ตารางแสดงการเปรียบเทียบงานวิจัยที่ศึกษา (ต่อ)

งานวิจัย	ข้อเด่น	ข้อด้อย	ข้อเสนอแนะในการปรับปรุง
Toward Semantic Web Information Extraction (Popov et al., 2003)	- มีการวิเคราะห์โครงสร้างไวยากรณ์ของประโยค และการวิเคราะห์คำที่อ้างอิงกันในเอกสาร - แยกส่วนของการจัดเก็บเอกสาร เมตาดาต้า และฐานความรู้ - ใช้ได้กับข้อมูลทุกโดเมน - ระบบสามารถ plug-in เข้า web browser ทำให้สามารถทำงานบน browser ได้ - มีการอธิบายรายละเอียดเพิ่มเติมของข้อมูลที่กำกับจากฐานความรู้ที่จัดเก็บ	- ต้องสร้างออนโทโลจีขนาดใหญ่เพื่อให้ครอบคลุมกับข้อมูลทั่ว ๆ ไป - ระบบไม่สามารถระบุความสัมพันธ์ระหว่างเอกสารได้	เนื่องจากในการกำกับข้อมูลในงานวิจัยนี้จะเป็นการกำกับข้อมูลทั่ว ๆ ไป ที่สามารถใช้ได้กับทุกโดเมน ซึ่งถ้าต้องการข้อมูลที่มีความเฉพาะของแต่ละโดเมนควรปรับปรุงในส่วนของออนโทโลจี โดยการสร้างออนโทโลจีเฉพาะของแต่ละโดเมน เพื่อให้ครอบคลุมกับเนื้อหาในแต่ละโดเมน ซึ่งในงานวิจัยนี้จะให้ความสนใจไปที่ความหมายของแต่ละคำหรือกลุ่มคำ โดยไม่ได้พิจารณาความสัมพันธ์ระหว่างกลุ่มคำนั้นที่ปรากฏในเอกสาร ดังนั้นควรพิจารณาในด้านความสัมพันธ์นี้ด้วย

ตารางที่ 3.2 ตารางแสดงการเปรียบเทียบงานวิจัยที่ศึกษา (ต่อ)

งานวิจัย	ข้อเด่น	ข้อด้อย	ข้อเสนอแนะในการปรับปรุง
Automatic Ontology-Based Knowledge Extraction from Web Documents (Alani et al., 2002)	- มีการวิเคราะห์โครงสร้างไวยากรณ์ของประโยค - มีการนำ wordnet ช่วยในการวิเคราะห์ความหมายของคำในประโยค - ระบุความสัมพันธ์ของข้อมูลในประโยคได้ - โดเมนออนโทโลจีมีการปรับปรุงอัตโนมัติ โดยการเพิ่มอินสแตนซ์ที่สอดคล้องกับคลาส	- ในการวิเคราะห์ความสัมพันธ์ในประโยคโดยใช้ออนโทโลจีจะสามารถทำการวิเคราะห์ความสัมพันธ์ได้เพียง 2 สัมพันธ์สำหรับคำกริยา 1 คำเท่านั้น	ในงานวิจัยนี้ได้พิจารณาสัมพันธ์ระหว่างคำโดยอาศัยความสัมพันธ์จากออนโทโลจีที่กำหนดนำมาใช้ในการวิเคราะห์ความสัมพันธ์ระหว่างกลุ่มคำในประโยค ซึ่งความสัมพันธ์ที่ได้นั้นเป็นความสัมพันธ์ที่ไม่ซับซ้อน ควรนำเทคนิคอื่นที่สามารถทำให้ปรากฏความสัมพันธ์ภายในประโยคเพิ่มมากขึ้นกว่าเดิม เช่นการใช้ parser ในการวิเคราะห์ความสัมพันธ์
ONTEA : Platform for pattern based automated semantic annotation (Laclavik et al., 2009)	- โดเมนออนโทโลจีมีการปรับปรุงอัตโนมัติ โดยการเพิ่มอินสแตนซ์ที่สอดคล้องกับคลาส - มีการแยกส่วนของการจัดเก็บแพทเทิร์นและโดเมนออนโทโลจี - ระบบสามารถระบุความสัมพันธ์ระหว่างข้อมูลในเอกสารได้ผ่านทางออนโทโลจี	- การวิเคราะห์ข้อมูลให้สอดคล้องกับคอนเซ็ปต์ต้องอาศัย regular expression patterns เพียงอย่างเดียว - ไม่มีการวิเคราะห์โครงสร้างของประโยค	ในการกำกับความหมายให้กับกลุ่มคำในเอกสารนั้นจะอาศัยแพทเทิร์น และกลุ่มคำศัพท์ที่เกี่ยวข้องที่กำหนดไว้ในออนโทโลจีเท่านั้น โดยไม่ได้พิจารณาโครงสร้างประโยคของข้อมูลที่กำกับ ซึ่งในการกำกับนั้นสนใจแต่เฉพาะกลุ่มคำ ดังนั้นควรเพิ่มการวิเคราะห์โครงสร้างประโยคที่ใช้ในการกำกับความหมายด้วย

ตารางที่ 3.2 ตารางแสดงการเปรียบเทียบงานวิจัยที่ศึกษา (ต่อ)

งานวิจัย	ข้อเด่น	ข้อด้อย	ข้อเสนอแนะในการปรับปรุง
Semi-Automatic Approach for Semantic Annotation (Yasrebi and Mohsenadeh, 2009)	- มีการใช้ wordnet ช่วยในการวิเคราะห์ข้อมูล เมื่อไม่พบข้อมูลที่ต้องการกำกับในฐานความรู้ เพื่อดูว่าข้อมูลนั้นสอดคล้องกับข้อมูลใดในฐานความรู้	- ไม่มีการวิเคราะห์โครงสร้างของประโยคในการกำกับข้อมูล - ต้องใช้ฐานความรู้ขนาดใหญ่ในการกำกับข้อมูล	การระบุความหมายของกลุ่มคำในเอกสารจะใช้ฐานความรู้ที่กำหนดความหมายของแต่ละกลุ่มคำภายในโดเมน และถ้าคำนั้นไม่ปรากฏในฐานความรู้ก็จะนำไปค้นหาใน wordnet ว่ามีความหมายอยู่ในกลุ่มใดได้บ้าง ซึ่งถ้าคำในเอกสารปรากฏในฐานความรู้เป็นจำนวนน้อยแล้ว จะต้องนำคำที่ไม่ปรากฏทั้งหมดไปค้นหาใน wordnet แล้วทำการวิเคราะห์เพื่อระบุว่าคำนั้นควรมีความหมายอยู่ในกลุ่มใด เช่น sport state building เป็นต้น ซึ่งคำเฉพาะบางครั้งก็ไม่ปรากฏในข้อมูล wordnet ดังนั้นถ้ามีการใช้แพทเทิร์น หรือคำสำคัญมาช่วยในการระบุกลุ่มความหมาย และใช้เทคนิค name entity recognition มาช่วยในการสกัดชื่อหน่วยงาน ชื่อบุคคลและชื่อสถานที่

ตารางที่ 3.2 ตารางแสดงการเปรียบเทียบงานวิจัยที่ศึกษา (ต่อ)

งานวิจัย	ข้อเด่น	ข้อด้อย	ข้อเสนอแนะในการปรับปรุง
Weather Forecast Data Semantic Analysis in F-logic (Mestrovic et al., 2007)	- แบ่งกลุ่มข้อมูลของข้อความก่อน ทำให้สามารถเลือกใช้เทมเพลตได้เหมาะสมกับข้อมูล เช่น weather forecast, sea weather forecast, temperature เป็นต้น - มีการวิเคราะห์ความสัมพันธ์ระหว่างประโยคย่อย ในประโยคที่ซับซ้อน ทำให้สามารถใช้ข้อมูลร่วมกันได้ระหว่างประโยคย่อยได้ - รูปแบบในการจัดเก็บผลลัพธ์สามารถนำไปใช้ประโยชน์ต่อได้ง่าย เพราะว่ามีการจัดเก็บในรูปแบบของ เฟรมสล๊อต	- ต้องสร้างข้อมูลความรู้ให้กับระบบเป็นจำนวนมากเพื่อให้ครอบคลุมกับข้อมูลที่ต้องการวิเคราะห์	จากงานวิจัยนี้การวิเคราะห์ความหมายของกลุ่มคำนั้นจะใช้ฐานความรู้ที่กำหนดขึ้น นั่นก็คือดิกชันนารี เพียงอย่างเดียว ซึ่งดิกชันนารีนั้นจะต้องครอบคลุมข้อมูลที่เกี่ยวกับโดเมนสภาพอากาศทั้งหมด ดังนั้นเพื่อเป็นการลดจำนวนข้อมูลในดิกชันนารี อาจใช้เทคนิค name entity recognition ในการกำหนดข้อมูลสถานที่ หรือใช้ regular expression patterns ในการกำหนดข้อมูลที่มีรูปแบบเฉพาะ เป็นต้น

ตารางที่ 3.2 ตารางแสดงการเปรียบเทียบงานวิจัยที่ศึกษา (ต่อ)

งานวิจัย	ข้อเด่น	ข้อด้อย	ข้อเสนอแนะในการปรับปรุง
Cerno : Light-weight tool support for semantic annotation of textual documents (Kiyavitskaya et al., 2009)	- ไม่จำเป็นต้องใช้ฐานความรู้ขนาดใหญ่ เช่น การใช้ออนโทโลจี เพียงแต่สร้างรายการ indicator ของแต่ละคอนเซ็ปที่ต้องการกำกับความหมาย	- ไม่มีการวิเคราะห์โครงสร้างไวยากรณ์และความหมายของประโยค	เนื่องจากในงานวิจัยนี้ใช้กฎในการวิเคราะห์ความหมายของกลุ่มคำ เช่นระบุข้อมูลราคา วันที่ อีเมล์ เป็นต้น และใช้คำสำคัญในการระบุความหมายของประโยค เช่น ข้อมูลสิ่งอำนวยความสะดวกของสถานที่พัก ซึ่งถึงแม้จะทำงานได้รวดเร็วกว่าการใช้ออนโทโลจีเป็นฐานความรู้ แต่อย่างไรก็ตามออนโทโลจีก็ยังสามารถอธิบายข้อมูลได้ละเอียดและสามารถระบุความสัมพันธ์ระหว่างข้อมูลได้

จากงานวิจัยข้างต้น พบว่าใน (Alani et al., 2002, Popov et al., 2003, and Laclavik et al., 2009) ใช้ออนโทโลจีเป็นรูปแบบการจัดเก็บความรู้อย่างหนึ่งที่มีการจัดเก็บข้อมูลในลักษณะคอนเซ็ปและความสัมพันธ์ระหว่างคอนเซ็ป ทำให้สามารถจัดการข้อมูลได้อย่างยืดหยุ่น อีกทั้งเป็นการกำหนดข้อมูลอย่างมีระเบียบแบบแผน และสามารถแลกเปลี่ยนข้อมูลกันได้อย่างสะดวก ทำให้การใช้ออนโทโลจีทำหน้าที่ในการแทนความรู้ของระบบจึงเป็นวิธีการหนึ่งที่น่าสนใจ แทนการจัดเก็บความรู้ในรูปแบบฐานข้อมูลหรือไฟล์เอกสาร

การวิเคราะห์กลุ่มคำเพื่อกำกับความหมายนั้น ใน (Laclavik et al., 2009 and Tellez-Valero et al., 2009) ใช้ regular expression patterns ในการระบุกลุ่มคำ โดยกลุ่มคำที่ถูกระบุนั้นจะต้องมีรูปแบบที่แน่นอน เช่น วันที่ เวลา ข้อมูลหน่วยวัด เป็นต้น ส่วนของข้อมูลอื่นที่เกี่ยวข้องที่ต้องการกำกับความหมายนั้นจะใช้ฐานความรู้ที่กำหนดขึ้นได้แก่ ออนโทโลจี (Alani et al., 2002, Popov et al., 2003, and Laclavik et al., 2009) คีย์เวิร์ดของแต่ละคอนเซ็ป (Kiyavitskaya et al., 2009) ซึ่งการแทนความรู้ด้วยออนโทโลยีจะมีความยืดหยุ่นมากกว่า

อย่างไรก็ตามในการวิเคราะห์ข้อความเว็บเอกสารนั้น การวิเคราะห์ในระดับประโยคจะให้รายละเอียดของข้อความที่ละเอียดและมีความชัดเจน ซึ่งการวิเคราะห์โครงสร้างและความหมายของประโยคจึงเป็นสิ่งสำคัญสำหรับงานทางด้านการวิเคราะห์ข้อความ เพื่อให้เข้าใจความหมายของเนื้อหาของเอกสารก่อนทำการวิเคราะห์ข้อความ

ดังนั้นแนวทางการวิเคราะห์ข้อความเว็บเอกสารที่จะนำเสนอในต่อไปในบทที่ 4 จึงได้นำออนโทโลจีและเทคนิคทางด้านการประมวลผลภาษาธรรมชาติมาวิเคราะห์ข้อความเอกสาร และจัดเก็บผลลัพธ์จากการวิเคราะห์ไว้ในรูปแบบเฟรมสล็อต

บทที่ 4

การออกแบบระบบ

แนวทางการนำเสนอวิธีการการวิเคราะห์ข้อความเว็บเอกสารนี้ ได้มีการนำออนโทโลจี (Ontology) มาช่วยในการวิเคราะห์คำหรือกลุ่มคำในเชิงความหมาย และรูปแบบของเฟรมที่ใช้สำหรับแสดงผลลัพธ์ของการวิเคราะห์ข้อความ ร่วมกับเทคนิคในการประมวลผลภาษาธรรมชาติ เพื่อให้ระบบเข้าใจรูปแบบโครงสร้างและความหมายของประโยคในข้อความ

4.1 การออกแบบกระบวนการทำงาน

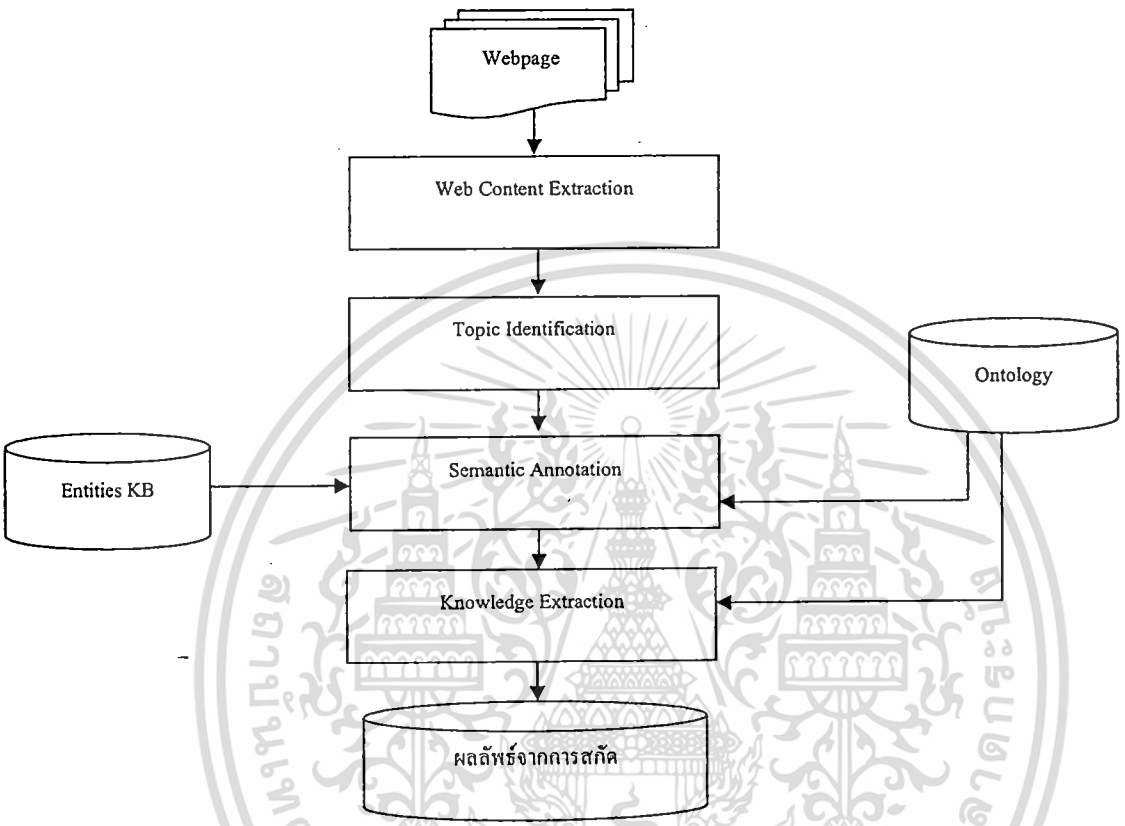
ในงานวิจัยนี้นำเสนอแนวทางการวิเคราะห์ข้อความจากเว็บเอกสาร โดยการใช้ออนโทโลจีร่วมกับเทคนิคการประมวลผลภาษาธรรมชาติเพื่อทำการวิเคราะห์โครงสร้างและความหมายของข้อความในระดับประโยค ข้อมูลนำเข้าของระบบคือเว็บเอกสารที่ต้องการวิเคราะห์ ออนโทโลจีของแต่ละโดเมนที่สนใจ และผลลัพธ์จากการวิเคราะห์ข้อความจะถูกจัดเก็บในรูปแบบของเฟรมสล็อต ซึ่งข้อมูลในแต่ละสล็อตจะอธิบายถึงองค์ประกอบต่าง ๆ ที่เราสนใจในโดเมนนั้น ๆ ในแต่ละโดเมนก็จะมีเฟรมที่แตกต่างกัน สำหรับในงานวิจัยนี้ นำเสนอเทคนิคที่วิเคราะห์ข้อความเอกสารเฉพาะโดเมน

ฐานความรู้ที่ใช้ในการวิเคราะห์ความหมายคือ ออนโทโลจี ซึ่งมีรูปแบบการจัดเก็บโครงสร้างข้อมูลที่เป็นระเบียบ และสามารถแสดงความสัมพันธ์ระหว่างข้อมูลได้ โดยสิ่งที่จัดเก็บในออนโทโลจีจะอยู่ในรูปแบบของคลาส อินสแตนซ์และความสัมพันธ์ โดยแต่ละโดเมนจะแทนด้วยออนโทโลจีที่แตกต่างกันขึ้นกับเนื้อหาของแต่ละโดเมน และฐานความรู้อีกชนิดหนึ่งได้แก่ ข้อมูลของคำนามเฉพาะ (Entities Knowledge Base) ได้แก่ ชื่อสถานที่ ชื่อหน่วยงานและบุคคล โดยฐานความรู้นี้จะทำการสกัดจากเอกสารตัวอย่างและจากเอกสารที่ใช้ในการวิเคราะห์ โดยใช้เทคนิคที่เรียกว่า Name Entity Recognition

กระบวนการทำงานของระบบการวิเคราะห์ข้อความเว็บเอกสารสามารถแบ่งออกเป็น 4 ขั้นตอนหลัก แสดงดังภาพที่ 4.1

1. Web Content Extraction ขั้นตอนของเตรียมข้อมูลสำหรับการประมวลผล โดยทำการลบแท็ค HTML ออกจากเอกสารเพื่อดึงเฉพาะข้อความของเอกสารเท่านั้น
2. Topic Identification กำหนดหัวข้อของเอกสารให้สอดคล้องกับข้อความในเอกสาร
3. Semantic Annotation กำกับบทบาททางความหมาย (Semantic Roles Annotation) ให้ของแต่ละกลุ่มคำหรือวลีในประโยค

4. Knowledge extraction สกัดความรู้จากเอกสาร ซึ่งในขั้นตอนนี้จะทำการพิจารณาโครงสร้างทางความหมายของประโยคด้วย ซึ่งวิเคราะห์ได้จากการใช้เทคนิคการแจงแจงไวยากรณ์เชิงความหมาย จากนั้นจึงทำการจัดเก็บข้อมูลที่วิเคราะห์ได้ในรูปแบบของเฟรม



ภาพที่ 4.1 สถาปัตยกรรมของแนวทางในการวิเคราะห์ข้อความเว็บเอกสาร

4.1.1 Web Content Extraction

ด้วยเอกสารที่ทำการวิเคราะห์นั้นมุ่งเน้นไปที่เว็บเอกสาร ซึ่งภายในเอกสารนั้นประกอบด้วยแท็ค HTML รูปภาพ โฆษณา มัลติมีเดีย ซึ่งไม่ใช่ข้อความที่เราต้องการวิเคราะห์ ส่วนของข้อความเท่านั้นที่ถูกใช้ในการวิเคราะห์ ดังนั้นจึงต้องทำการลบข้อมูลเหล่านี้ออก ให้เหลือเพียงส่วนของข้อความเท่านั้น

MEXICO CITY (Reuters) – A tropical depression formed in the Gulf of Mexico late on Wednesday and was set to slam into the Gulf Coast near the Texas-Mexico border on Thursday, a region still recovering from Hurricane Alex, the U.S. National Hurricane Center said in a report.

Another serious storm in the Gulf of Mexico could further disrupt efforts to contain BP's massive oil spill off the Louisiana coast.

A tropical storm warning was issued in the lower Rio Grande valley along the border, from south of Baffin Bay, Texas to Rio San Fernando, Mexico. The warning signaled the storm could make landfall within the next 24 hours.

It would be the second named storm of the Atlantic hurricane season after Hurricane Alex battered northern Mexico last week, dumping heavy rains and flooding the major Mexican city of Monterrey in the state of Nuevo Leon, killing 12 people.

Alex, a Category 2 storm when it hit, shuttered some oil and gas production in the Gulf of Mexico as a precaution and delayed BP's efforts to capture oil gushing from its damaged well.

The new depression was expected to drench the border with up to 10 inches of rain in some places and cause strong winds by Thursday afternoon or evening.

ภาพที่ 4.2 แสดงภาพการสกัดข้อความออกจากเว็บเอกสาร

4.1.2 Topic Identification

เนื่องจากข้อความในเอกสารแต่ละโดเมนก็มีความแตกต่างกัน ทำให้ความรู้ที่ใช้ในการวิเคราะห์แต่ละโดเมนก็มีความแตกต่างกันไปด้วย ดังนั้นจึงต้องมีการระบุหัวข้อของเอกสารก่อนว่าเป็นข้อความเกี่ยวกับเรื่องอะไร เช่น เป็นข้อมูลข่าวกีฬา ข่าวพยากรณ์อากาศ ข่าวเศรษฐกิจ ข่าวสุขภาพ เป็นต้น เทคนิคในการระบุหัวข้อนี้จะใช้วิธีการทางสถิติเพื่อความรวดเร็วในการประมวลผล โดยรวบรวมเอกสารที่เกี่ยวข้องในแต่ละโดเมน แล้ววิเคราะห์กลุ่มคำจากข้อมูลเอกสารตัวอย่างเพื่อหาเซตของกลุ่มคำที่สามารถใช้แยกเอกสารในแต่ละโดเมนได้อย่างมีประสิทธิภาพ เช่น ในโดเมนข่าวพยากรณ์อากาศมักจะปรากฏคำที่ที่อยู่ในกลุ่มของสภาพอากาศ และลักษณะของอากาศ เป็นต้น

4.1.3 Semantic Annotation

ในการกำกับความหมายให้กับข้อมูลทีละประโยคนั้น จะต้องคำนึงด้วยว่าประโยคนั้นเกี่ยวข้องกับโดเมนที่ระบุหรือไม่ หรือเป็นในส่วนของการขยายข้อความที่ต้องการอธิบายเท่านั้น ซึ่งไม่เป็นสาระสำคัญของข้อความในเอกสาร ทำให้แต่ละประโยคจะถูกวิเคราะห์รูปแบบโครงสร้างของประโยคก่อน โดยในประโยคหนึ่งนั้นอาจประกอบด้วยประโยคย่อย หรือส่วนขยายที่ทำหน้าที่เป็นอนุประโยคของประโยคหลัก และต้องพิจารณาว่าในประโยคย่อยแต่ละประโยคนั้นอธิบายถึงรายละเอียดหลักในโดเมนนั้นหรือไม่

เช่น โดเมนพยากรณ์อากาศ ประโยค : The storm was moving northwest near 10 mph, and a gradual turn to the west-northwest was expected over the next 48 hours. ประกอบด้วยประโยคย่อย 2 ประโยคได้แก่

- The storm was moving northwest near 10 mph

- a gradual turn to the west-northwest was expected over the next 48 hours

จะพบว่าประโยคย่อยทั้งสองนั้นเกี่ยวข้องกับข้อมูลหลักของสภาพอากาศ ซึ่งให้คำแนะนำโดยออนโทโลจีที่ทำหน้าที่เป็นฐานความรู้ในแต่ละโดเมน ทำให้ประโยคย่อยทั้งสองถูกกำกับความหมายในแต่ละกลุ่มคำหรือวลี และกำกับความหมายได้ดังนี้

ประโยค : Hurricane Bill spun northward toward New England Coast Saturday with wind and rain as official warned beach lovers to head indoor the night. ประกอบด้วยประโยคย่อย 2 ประโยคได้แก่

- Hurricane Bill spun northward toward New England Coast Saturday with wind and rain
- official warned beach lovers to head indoor the night

จากประโยคย่อยทั้งสองพบว่าในประโยคย่อยที่หนึ่งนั้นแสดงรายละเอียดของสภาพอากาศที่ปรากฏในออนโทโลจี แต่ในขณะที่ประโยคย่อยที่สองนั้นเป็นรายละเอียดของการเตือนภัยดังนั้นจึงเลือกกำกับความหมายเฉพาะในประโยคย่อยที่หนึ่งเท่านั้น

ในขั้นตอนของการกำกับความหมายให้กับแต่ละประโยคที่ทำการวิเคราะห์แล้วนั้น มีวิธีการกำกับความหมาย 3 วิธี ได้แก่

1. กำกับความหมายให้กับวลีโดยเทียบกับ Entities Knowledge Base แล้วกำกับความหมายตามชื่อของเอนติตี้ เช่น Location , Organization หรือ Person
2. กำกับความหมายให้กับวลีโดยใช้ Regular Expression Pattern ที่ระบุไว้ในส่วนของกำหนดความสัมพันธ์ในออนโทโลจี โดยกำกับความหมายตามชื่อของความสัมพันธ์ที่ปรากฏในออนโทโลจี
3. กำกับความหมายให้กับวลีโดยเปรียบเทียบกับข้อมูลระดับอินสแตนซ์ แล้วกำกับความหมายด้วยชื่อคอนเซ็ปของอินสแตนซ์นั้น

จากขั้นตอนข้างต้นในการกำกับความหมายจะได้

ประโยค : The storm was moving northwest near 10 mph, and a gradual turn to the west-northwest was expected over the next 48 hours. กำกับความหมายได้เป็น

และประโยค : Hurricane Bill spun northward toward New England Coast Saturday with wind and rain as official warned beach lovers to head indoor the night. กำกับความหมายได้เป็น

$[\text{Hurricane Bill}]_{\text{stormevent}}$ spun $[\text{northward}]_{\text{direction}}$ toward $[\text{New England Coast}]_{\text{coast}}$ $[\text{Saturday}]_{\text{dayofweek}}$ with $[[\text{wind}]_{\text{windevent}}$ and $[\text{rain}]_{\text{precipitationevent}}]$ as official warned beach lovers to head indoor the night

4.1.4 Knowledge Extraction

เมื่อกำกับความหมายให้กับแต่ละวลีในประโยคแล้ว จากนั้นก็ทำการวิเคราะห์โครงสร้างทางความหหมายของประโยคหล่านั้น โดยใช้เทคนิคของการแจกแจงประโยคเข้ามาช่วย เพื่อแสดงให้เห็นโครงสร้างบทบาทของกลุ่มคำที่ปรากฏในประโยค แล้วนำผลลัพธ์ที่ได้นั้นจัดเก็บในรูปแบบของเฟรม

การแจกแจงความหมายให้ประโยคนั้น จะใช้เทคนิคที่เรียกว่า Parsing โดยกำหนดกฎทางความหมายของการแจกแจงประโยคที่เรียกว่า Semantic Rules และกฎของคำศัพท์ เรียกว่า Lexical Rules ซึ่งกฎเหล่านี้จะให้คำแนะนำโดยออนโทโลจี เนื่องจากรูปแบบการจัดเก็บข้อมูลภายในออนโทโลจีเป็นลักษณะลำดับชั้น ซึ่งมีผลลัพธ์จากการแจกแจงความหมายก็มีลักษณะเป็นลำดับชั้นเช่นกัน ทำให้การเขียนกฎสามารถแนะนำได้โดยออนโทโลจี

ตัวอย่างการแจกแจงความหมายของประโยค : The storm was moving northwest near 10 mph, and a gradual turn to the west-northwest was expected over the next 48 hours.

Semantic Rules:

WeartherDomain -> WeatherEvent COMMA CC WeatherEvent

WeatherEvent -> Weather WeatherDirection | WeatherDirection WeatherDateTime

WeatherDirection -> direction windSpeed | direction

WeatherDateTime -> dateTime

Lexical Rules:

Weather -> “storm”

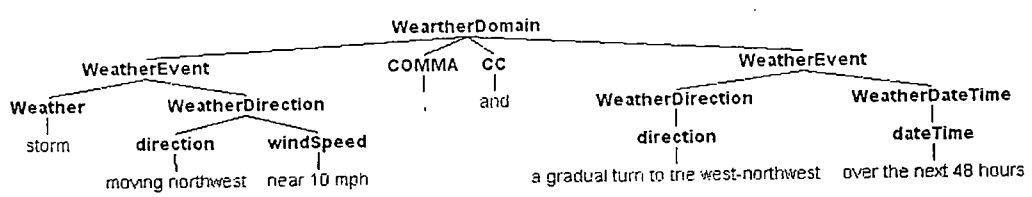
direction -> “moving northwest” | ”a gradual turn to the west-northwest ”

windSpeed -> “near 10 mph”

dateTime -> “over the next 48 hours”

COMMA -> “,”

CC -> “and”



ภาพที่ 4.3 โครงสร้างต้นไม้ความหมายของประโยค “The storm was moving northwest near 10 mph, and a gradual turn to the west-northwest was expected over the next 48 hours.”

จากการวิเคราะห์ในภาพที่ 4.3 พบว่ามีเหตุการณ์ของสภาพอากาศปรากฏอยู่สองเหตุการณ์ ในเหตุการณ์แรกเป็นข้อมูลของ Storm และในเหตุการณ์ที่สองนั้นไม่ปรากฏสภาพอากาศ แต่จากประโยคนั้นเป็นลักษณะคล้อยตามกัน ทำให้การจัดเก็บข้อมูลนั้นจะทำการปรับปรุงข้อมูลในการจัดเก็บในส่วนถัดไป

การจัดเก็บข้อมูลลงในเฟรมสล็อต

เฟรมสล็อตที่กำหนดในแต่ละโดเมน ก็คือส่วนที่ระบุว่าต้องการจัดเก็บข้อมูลอะไรจากข้อความ โดยจะสกัดข้อมูลมาจัดเก็บลงในแต่ละสล็อต โดยเฟรมนั้นจะถูกกำหนดมาจากโครงสร้างในออนโทโลจี ผลลัพธ์ที่ได้จากการแจกแจงไวยากรณ์ความหมายจะถูกนำมาจัดเก็บในรูปแบบของเฟรมสล็อต

ตัวอย่างประโยค : The storm was moving northwest near 10 mph, and a gradual turn to the west-northwest was expected over the next 48 hours.

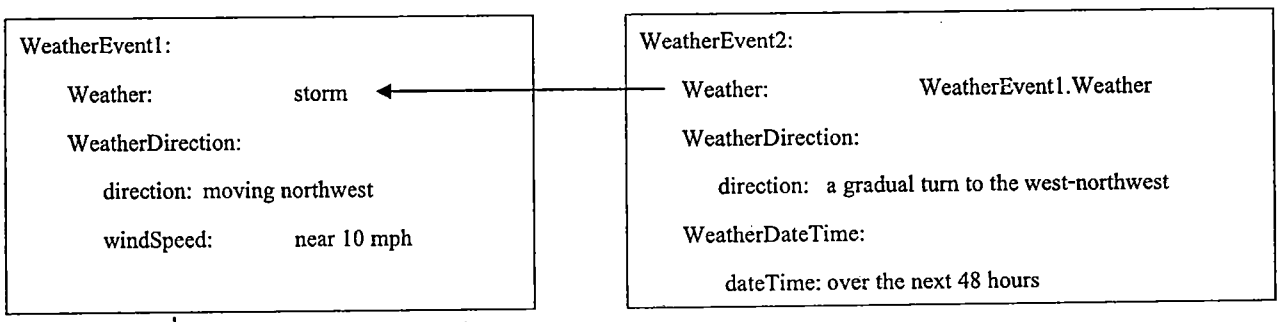
WeatherEvent1:
Weather: storm
WeatherDirection:
direction: moving northwest
windSpeed: near 10 mph

WeatherEvent2:
WeatherDirection:
Direction: a gradual turn to the west-northwest
WeatherDateTime:
dateTime: over the next 48 hours

จากประโยคข้างต้นพบว่าเฟรม WeatherEvent2 นั้นยังไม่สมบูรณ์เนื่องจากไม่ปรากฏข้อมูลสภาพอากาศ จึงปรับปรุงได้เป็น

WeatherEvent2:
Weather: WeatherEvent1
WeatherDirection:
direction: a gradual turn to the west-northwest
WeatherDateTime:
dateTime: over the next 48 hours

หรือสามารถเขียนได้อีกรูปแบบแสดงดังภาพที่ 4.4



ภาพที่ 4.4 ตัวอย่างเฟรมของประโยค "The storm was moving northwest near 10 mph, and a gradual turn to the west-northwest was expected over the next 48 hours."

4.2 การออกแบบหน้าจอระบบ

ระบบสามารถทำการวิเคราะห์เนื้อหาเว็บเอกสารทีละเว็บ แต่ผู้ใช้สามารถกำหนดเว็บเอกสารที่ต้องการวิเคราะห์ไว้ล่วงหน้าได้ โดยส่วนประกอบในหน้าจอระบบแสดงดังภาพที่ 4.5 นั้น ประกอบด้วย

ส่วนที่ 1 เมนูในการทำงานของระบบ ประกอบด้วย การกำหนดการทำงานเริ่มต้นของระบบ การเพิ่มเว็บเอกสาร การระบุหัวข้อเว็บเอกสาร การกำกับความหมาย การสกัดความรู้ และบันทึกผลลัพธ์ที่ได้จากการสกัดความรู้

ส่วนที่ 2 เป็นการจัดเก็บรายชื่อของเว็บเอกสารที่ต้องการวิเคราะห์ เมื่อต้องการวิเคราะห์เว็บเอกสารใดก็สามารถคลิกเลือกที่ชื่อเอกสารนั้น

ส่วนที่ 3 แสดงข้อความของเว็บเอกสารเท่านั้น โดยลบในส่วนของรูปภาพ โฆษณา ความคิดเห็นต่าง ๆ

ส่วนที่ 4 แสดงรายชื่อออนโทโลจีโดเมนที่รวบรวมไว้ในระบบ

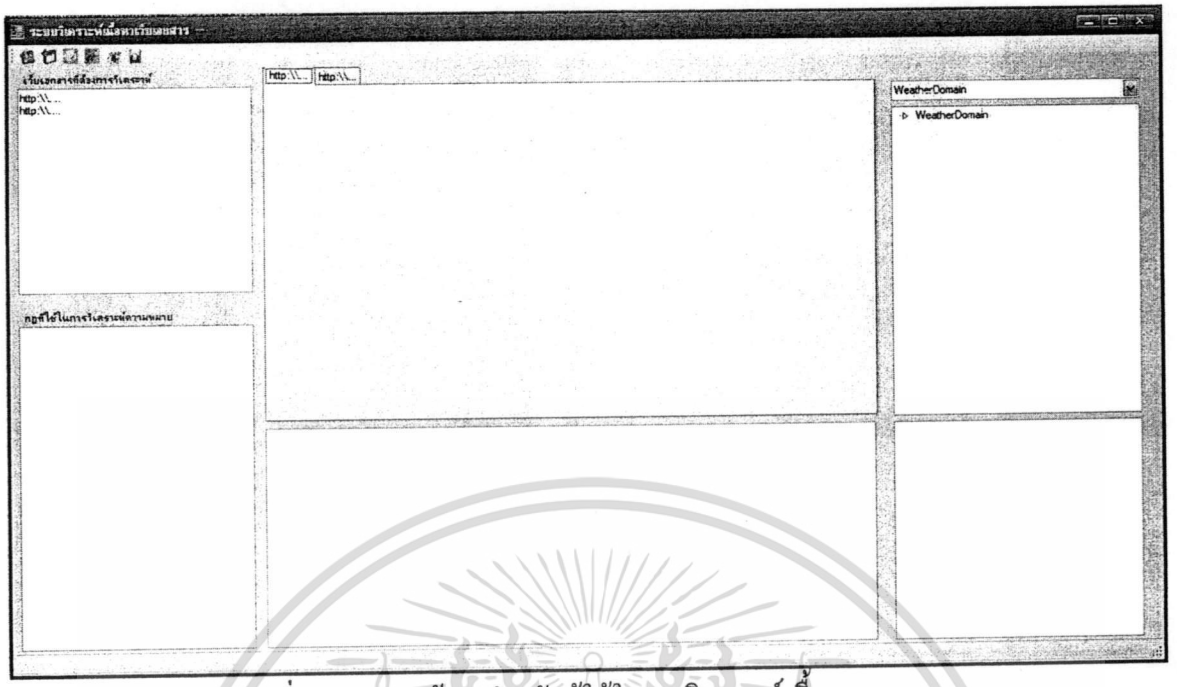
ส่วนที่ 5 แสดงออนโทโลจีที่สอดคล้องกับข้อความของเว็บเอกสารมากที่สุด

ส่วนที่ 6 แสดงกฎทั้งหมดที่ใช้ในการวิเคราะห์เชิงความหมายด้วย Parser ของเอกสารที่ทำการวิเคราะห์

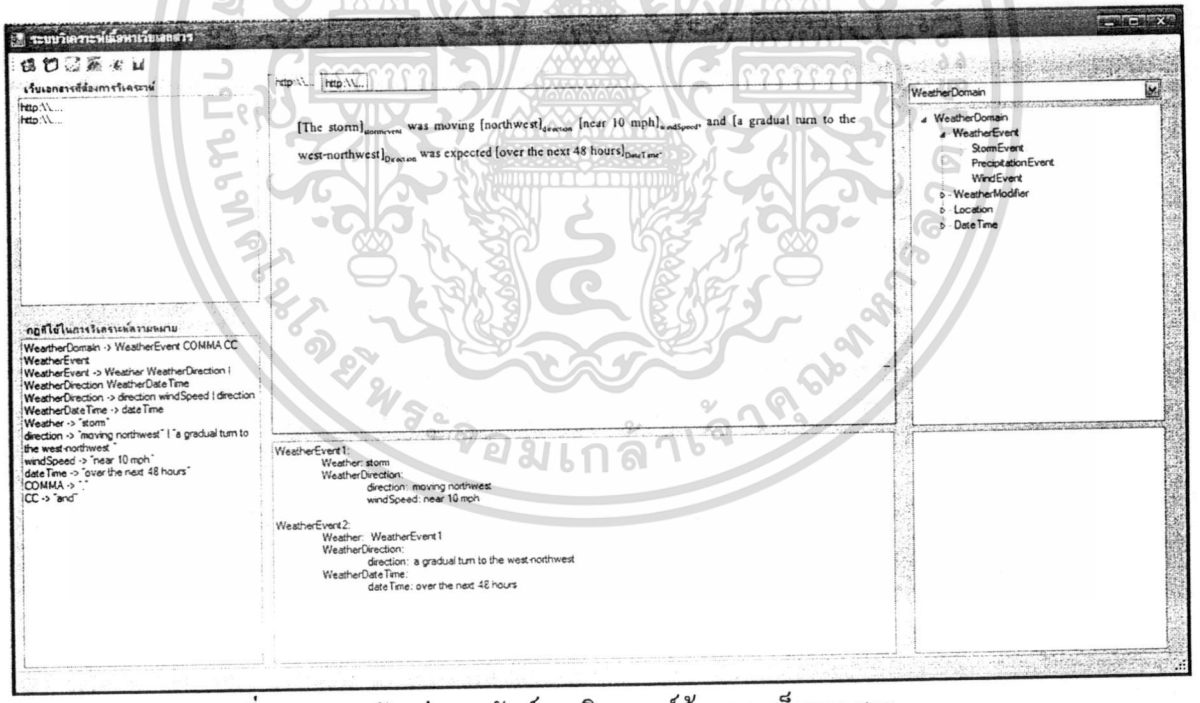
ส่วนที่ 7 แสดงผลลัพธ์จากการวิเคราะห์ข้อความเอกสารนั้น

ส่วนที่ 8 แสดงหมายเหตุ หรือ comment จากระบบ

ในภาพที่ 4.6 แสดงผลลัพธ์การทำงานของการวิเคราะห์ข้อความเอกสาร ซึ่งผลลัพธ์ที่ได้จากการวิเคราะห์นั้นสามารถจัดเก็บไว้เพื่อใช้งานอื่นได้



ภาพที่ 4.5 แสดงหน้าจอสำหรับผู้ใช้ในการวิเคราะห์เนื้อหา



ภาพที่ 4.6 แสดงตัวอย่างผลลัพธ์การวิเคราะห์ข้อความเว็บเอกสาร

4.3 เปรียบเทียบงานที่นำเสนอกับงานวิจัยอื่น ๆ

แนวทางที่นำเสนอนี้เป็นการนำแนวคิดในงานวิจัยที่ได้นำเสนอในบทที่ 2 มาปรับปรุงและเพิ่มแนวคิดใหม่ โดยงานวิจัยที่ใกล้เคียงกับแนวคิดที่นำเสนอใหม่นี้ประกอบด้วย 5 งานวิจัยได้แก่

งานวิจัย 1: Using Machine Learning for Extracting Information from Natural Disaster News Reports (Tellez-Valero et al., 2009)

งานวิจัย 2: Toward Semantic Web Information Extraction (Popov et al., 2003)

งานวิจัย 3: Automatic Ontology-Based Knowledge Extraction from Web Documents (Alani et al., 2002)

งานวิจัย 4: ONTEA : Platform for pattern based automated semantic annotation (Laclavik et al., 2009)

งานวิจัย 5: Weather Forecast Data Semantic Analysis in F-logic (Mestrovic et al., 2007)

ซึ่งได้ทำการเปรียบเทียบเทคนิคที่ใช้ในแนวคิดใหม่นี้กับแนวคิดในงานวิจัยข้างต้น โดยแสดงในตารางที่ 4.1

ตารางที่ 4.1 แสดงรายละเอียดการเปรียบเทียบการทำงานระหว่างแนวทางที่นำเสนอในเอกสารฉบับนี้กับงานวิจัยที่มีแนวคิดใกล้เคียง

รายละเอียด	แนวทางที่นำเสนอ	งานวิจัย 1	งานวิจัย 2	งานวิจัย 3	งานวิจัย 4	งานวิจัย 5
1. ใช้ออนโทโลจีเป็นรูปแบบในการจัดเก็บความรู้	√		√	√	√	
2. ทำงานกับข้อมูลในโดเมนเฉพาะ	√	√		√	√	√
3. ใช้เทคนิค name entity recognition ในการวิเคราะห์ชื่อสถานที่ หน่วยงาน และบุคคล	√		√			
4. มีการคัดเลือกเอกสารที่เกี่ยวข้องกับโดเมน	√	√		√		
5. จัดกลุ่มเอกสารให้สอดคล้องกับโดเมนของเอกสาร	√					√
6. ใช้ regular expression patterns ในการวิเคราะห์กลุ่มคำเพื่อกำกับความหมาย	√	√			√	
7. กำกับความหมายให้กับกลุ่มคำที่เป็นวลี	√					
8. วิเคราะห์ความสัมพันธ์ทางความหมายใน	√			√		

ระดับประโยค โดยใช้ความสัมพันธ์ที่กำหนดขึ้นในออนโทโลจี						
9. ใช้เทคนิคการแจกแจงความหมาย (parser) ในการวิเคราะห์ความสัมพันธ์ระหว่างกลุ่มคำในประโยค	√					
10. จัดเก็บข้อมูลที่ทำการวิเคราะห์แล้วในรูปแบบเฟรมสล็อต	√					√
11. จัดเก็บข้อมูลที่ทำการวิเคราะห์แล้วเป็นอินสแตนซ์ในโดเมนออนโทโลจี			√	√	√	

จากตารางที่ 4.1 พบว่าแนวทางที่นำเสนอนั้นมีความคล้ายคลึงกับงานวิจัยที่ 3 ค่อนข้างมาก แต่อย่างไรก็ตาม ก็มีความคล้ายคลึงในบางส่วนกับงานวิจัยอื่น ๆ ซึ่งจะอธิบายรายละเอียดดังต่อไปนี้

แนวทางที่นำเสนอนี้จะมีความคล้ายคลึงกับงานวิจัยที่ 1 คือมีการคัดเลือกเอกสารที่เกี่ยวข้องกับโดเมนของระบบ ในงานวิจัยที่ 1 นี้จะเป็นการคัดกรองเอกสารที่เกี่ยวข้องออกจากเอกสารที่ไม่เกี่ยวข้องกับโดเมนโดยอาศัยเทคนิคการแจกแจงเอกสารและมีการกำหนดแพทเทิร์นด้วย regular expression เพื่อระบุความหมายของกลุ่มคำที่เกี่ยวข้องกับโดเมน ซึ่งในแนวทางที่นำเสนอนี้ก็มีการคัดเลือกเอกสารที่เกี่ยวข้องก่อนในขั้นแรก แต่เมื่อคัดเลือกเอกสารแล้วยังได้ระบุกลุ่มของเอกสารนั้นด้วย อีกทั้งยังสามารถระบุเอกสารนั้นไม่อยู่ภายในกลุ่มของโดเมนที่กำหนดด้วย

แนวทางที่นำเสนอนี้จะมีความคล้ายคลึงกับงานวิจัยที่ 2 คือ มีการใช้ออนโทโลจีในการจัดเก็บความรู้ที่ใช้ในการวิเคราะห์ และใช้เทคนิค name entity recognition ในการระบุชื่อหน่วยงาน ชื่อบุคคล และชื่อสถานที่ แต่อย่างไรก็ตามในแนวทางที่นำเสนอนี้ได้มีวิเคราะห์กลุ่มคำที่จะนำมากำกับก่อน โดยวิเคราะห์ว่าประโยคนั้นเกี่ยวข้องกับโดเมนนั้นหรือไม่ แล้วจึงนำประโยคนั้นมาวิเคราะห์กลุ่มคำเพื่อกำกับความหมาย ทำให้คำเดียวกันแต่ปรากฏในประโยคที่มีความหมายแตกต่างกันอาจจะถูกกำกับความหมายหรือไม่ก็ได้

แนวทางที่นำเสนอนี้จะมีความคล้ายคลึงกับงานวิจัยที่ 4 จะใช้แนวคิดในการใช้ regular expression patterns และฐานความรู้ของคำศัพท์ที่กำหนดในออนโทโลจีมาระบุกลุ่มคำที่เกี่ยวข้องกับโดเมน แต่อย่างไรก็ตามในงานวิจัยนี้เอกสารที่ทำการกำกับความหมายจะต้องถูกเลือกว่าเป็นเอกสารที่ถูกต้องแล้วจึงทำการกำกับความหมาย และไม่มีการวิเคราะห์กลุ่มคำก่อนที่จะนำมากำกับความหมาย

แนวทางที่นำเสนอนี้จะมีความคล้ายคลึงกับงานวิจัยที่ 5 เป็นงานวิจัยที่ใช้กลุ่มโดเมนเหมือนกัน คือข่าวพยากรณ์อากาศและมีการจัดเก็บในรูปแบบเฟรมสล็อต แต่วิธีการวิเคราะห์

ข้อมูลนั้นมีความแตกต่างกัน โดยในงานวิจัยนี้จะเก็บข้อมูลความรู้ของคำศัพท์ที่เกี่ยวข้องกับโดเมนในรูปแบบของดิกชันนารี แต่แนวทางที่นำเสนอนี้จะจัดเก็บฐานความรู้ไว้ในออนโทโลจีซึ่งมีความยืดหยุ่นและสามารถระบุความสัมพันธ์ระหว่างกลุ่มคำเหล่านั้นได้

และงานวิจัยที่คล้ายคลึงกับแนวคิดที่นำเสนอคืองานวิจัยที่ 3 เนื่องจากมีการใช้ออนโทโลจีเป็นฐานความรู้ในการวิเคราะห์ความหมายของกลุ่มคำและมีการวิเคราะห์ความสัมพันธ์เชิงความหมายระหว่างกลุ่มคำในประโยค โดยในงานวิจัยนี้จะระบุความสัมพันธ์นี้จากความสัมพันธ์ที่กำหนดไว้ในออนโทโลจี ซึ่งจะมีข้อจำกัดว่า ความสัมพันธ์ระหว่างกลุ่มคำจะมีได้เพียง 2 ความสัมพันธ์เช่น ศิลปินเกิดวันที่ ณ สถานที่ แต่ถ้าต้องการระบุว่าศิลปินเรียนจบปริญญาตรีจากมหาวิทยาลัยอะไร เมื่อปีค.ศ. ใดนั้นจะไม่สามารถกำหนดได้ ดังนั้นในงานวิจัยนี้จึงได้มีการใช้ parser แทนในการระบุความสัมพันธ์ระหว่างกลุ่มคำในประโยค และจัดเก็บข้อมูลในรูปแบบของเฟรมสล็อต

บทที่ 5
สรุปและแนวทางในอนาคต

5.1 สรุปงานวิจัย

ในงานวิจัยนี้เป็นการศึกษาและเปรียบเทียบเทคนิคในการวิเคราะห์ข้อความเว็บเอกสาร พร้อมทั้งนำเสนอแนวทางในการวิเคราะห์ข้อความเว็บเอกสาร โดยแบ่งเทคนิคทางด้านการวิเคราะห์ข้อความเอกสารออกเป็น 4 กลุ่มด้วยกัน คือการวิเคราะห์ข้อความด้วยวิธีการเปรียบเทียบรูปแบบ (Pattern-Matching) การวิเคราะห์ข้อความด้วยวิธีการเรียนรู้ของเครื่องจักร (Machine Learning) การวิเคราะห์ข้อความโดยการใช้ออนโทโลจี และการวิเคราะห์โดยการใช้ฐานความรู้อื่นและนิยมทำงานกับโดเมนเฉพาะเรื่อง ซึ่งมีขอบเขตของรายละเอียดในเนื้อหาของเรื่องนั้นที่ชัดเจนและจำกัด แต่อย่างไรก็ตามในการที่จะทำให้สามารถวิเคราะห์ข้อความเอกสารได้กับทุกโดเมนก็สามารถทำได้ เนื่องจากในการวิเคราะห์เนื้อต้องอาศัยฐานความรู้ในการให้ความรู้กับระบบเพื่อให้เข้าใจความหมายของคำที่ปรากฏในเอกสารและระบุผลลัพธ์จากการวิเคราะห์ว่าต้องการวิเคราะห์ในเนื้อหาอะไรบ้างนั้น เมื่อเราเพิ่มความรู้ต่าง ๆ เหล่านี้ให้กับระบบมากขึ้น ระบบนั้นก็สามารถขยายให้สามารถทำงานได้กับทุกโดเมน ดังนั้นความรู้สำหรับระบบจึงเป็นสิ่งสำคัญสำหรับงานทางด้านการวิเคราะห์ข้อความเว็บเอกสาร

ออนโทโลจีเป็นรูปแบบการแทนความรู้แบบหนึ่งที่เป็นที่นิยมใช้กันมากในปัจจุบัน โดยเป็นลักษณะการอธิบายขอบเขตของข้อมูลในแต่ละโดเมน เนื่องจากคุณสมบัติและลักษณะในการอธิบายข้อมูลนั้นเป็นโครงสร้างแบบลำดับชั้น และสามารถสร้างความสัมพันธ์ระหว่างข้อมูลนั้นได้ ทำให้เกิดการเชื่อมโยงและแลกเปลี่ยนระหว่างของข้อมูลได้ อีกทั้งยังสามารถแชร์ความรู้ระหว่างออนโทโลจีได้ ทำให้จึงเป็นที่นิยม แต่อย่างไรก็ตามการสร้างออนโทโลจีนั้นมนุษย์จะเป็นผู้สร้างขึ้น เพื่อที่ว่าความรู้ที่ได้จากการวิเคราะห์จะมีความถูกต้องสูงกว่าการสร้างออนโทโลจีแบบอัตโนมัติ และตรงตามประเด็นกับข้อความที่ต้องการมากกว่า ซึ่งนับว่าเป็นข้อจำกัดอย่างหนึ่ง

การวิเคราะห์ข้อความเว็บเอกสารนอกจากการใช้ฐานความรู้เป็นข้อมูลในการวิเคราะห์แล้ว การใช้เทคนิคทางด้านการประมวลผลภาษาธรรมชาติ จะทำให้เข้าใจโครงสร้างทางไวยากรณ์และโครงสร้างทางความหมายของประโยคดียิ่งขึ้น ซึ่งมีส่วนสำคัญอย่างยิ่งในงานทางด้านการวิเคราะห์ข้อความเว็บเอกสาร เทคนิคทางด้านการประมวลผลภาษาธรรมชาติที่ใช้ในการทำงานทางด้านนี้ ได้แก่ การวิเคราะห์โครงสร้างไวยากรณ์ด้วย Parser เพื่อวิเคราะห์กลุ่มคำที่ทำหน้าที่เป็นวลี การวิเคราะห์อนุประโยค การอ้างอิงของคำที่ใช้แทนกันในประโยค เป็นต้น

ดังนั้นในงานวิจัยนี้จึงได้นำเสนอแนวทางในการวิเคราะห์เว็บเอกสาร ที่มีการใช้ออนโทโลจีและเทคนิคการประมวลผลภาษาธรรมชาติ โดยแนวทางนั้นได้มีการเพิ่มการใช้เทคนิคการ

แจกแจงความหมายของประโยคเข้ามาช่วยในการวิเคราะห์โครงสร้างทางความหมายเพื่อให้การวิเคราะห์ข้อความนั้นมีความชัดเจนและถูกต้องมากยิ่งขึ้น โดยแนวทางที่นำเสนอนั้นจะทำการวิเคราะห์ในระดับประโยคของข้อความในเอกสาร แล้วทำการกำกับความหมายให้กับกลุ่มคำหรือวลีในประโยค จากนั้นจึงทำการวิเคราะห์โครงสร้างทางความหมายของประโยค เพื่อนำข้อมูลที่วิเคราะห์ได้จัดเก็บในรูปแบบเฟรมสล็อต ซึ่งเป็นรูปแบบในการแทนความรู้อย่างหนึ่ง

5.2 แนวทางในอนาคต

1. พัฒนาฐานความรู้ในแต่ละโดเมนเพื่อกำหนดขอบเขตของข้อมูลและผลลัพธ์ที่ต้องการจากการวิเคราะห์ข้อความ
2. พัฒนาและทดสอบระบบตามแนวทางที่นำเสนอในงานวิจัยนี้

บรรณานุกรม

Riloff E. and Schmelzenbach M. "An Empirical Approach to Conceptual Case Frame Acquisition," *Proceedings of the Sixth Workshop on Very Large Corpora (WVLC-98)*, 1998.

Tellez-Valero A., Montes-y-Gomez M., and Villasenor-Pineda L.," Using Machine Learning for Extracting Information from Natural Disaster News Reports," *Computation System,* Vol. 13 No. 1, 2009.

Popov B., Kiryakov A., Ognyanoff D., Manov D., Kirilov A., and Goranov M., "Towards Semantic Web Information Extraction," *Human Language Technologies Workshop at the 2nd International Semantic Web Conference (ISWC2003)*, 20 October 2003, Florida, USA.

Alani H., Kim S., Millard D. E., Weal J., Hall W., Lewis P. H., and Shadbolt N. R., "Automatic Ontology-Based Knowledge Extraction from Web Documents," *IEEE Intelligent Systems*, 18(1), 2003. pp. 14-21.

Laclavik M., Seleng M., CiGlan M., and Hluchy L., "ONTEA : Platform for pattern based automated semantic annotation," *Computation and Informations*, Vol. 28, 2009.

Mestrovic A., Martincic S., and Cubrilo M., "Weather Forecast Data Semantic Analysis in F-logic," *Journal of Information and Organization Sciences*, Vol 31 No 1, 2007.

Kiyavitskaya N., Zeni N., Cordy J. R., Mich L., and Mylopoulos J., "Cerno: Light-weight tool support for semantic annotation of textual documents," *Data & Knowledge Engineering* Volumn 68, Issue 12, December 2009, Pages 1470,1492.

Moschitti A., Morarescu P., and Harabagiu S. M., "Open Domain Information Extraction via Automatic Semantic Labeling," *In Proceeding of FLAIRS 2003*, pages 397-401, St. Augustine, FL.

M. Pasca, "Outclassing Wikipedia in Open-Domain Information Extraction: Weakly-Supervised Acquisition of Attributes over Conceptual Hierarchies," *EACL 2009*, pages 639-647.

Banko M., Cafarella M. J., Soderland S., Broadhead M., and Etzioni O., "Open Information Extraction from the Web," *In Proceedings of the International Joint Conference on Artificial Inteiligent*, 2007.

Ole R. Holsti R. O., "Content Analysis for the Social Sciences and Humanities," Reading, MA: Addison-Wesley. 1969

Kimberly A. N., "The Content Analysis Guidebook Online," [Online]. Available : http://academic.csuohio.edu/kneuendorf/content/, 2002.

Information Sciences and Interaction Sciences

The 3rd International Conference on Information Sciences and Interaction Sciences (ICIS 2010)

Chengdu, China , 23-25 June 2010

Yi Peng, Gang Kou, Franz I.S. Ko, Yong Zeng, Kae Dal Kwack (Eds.)

IEEE

IEEE computer society

AICIT

IEEE Conference Record Number : #16901
IEEE PDF Files Catalogue number: CFP1012K-ART
IEEE PDF files ISBN: 978-1-4244-7386-1
IEEE Print version Catalogue number: CFP1012K-PRT
IEEE Print version ISBN: 978-1-4244-7385-4

Additional copies may be ordered from:

IEEE Computer Society
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-131
Tel: + 1 800 272 6657
Fax: + 1 714 821 4641
http://computer.org/cspress
csbooks@computer.org

IEEE Service Center
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: + 1 732 981 0060
Fax: + 1 732 981 9667
http://shop.ieee.org/store/
customer-service@ieee.org

IEEE Computer Society
Asia/Pacific Office
Watanabe Bldg., 1-4-2
Minami-Aoyama
Minato-ku, Tokyo 107-006
JAPAN
Tel: + 81 3 3408 3118
Fax: + 81 3 3408 3553
Tokyo.ofc@computer.org

Individual paper REPRINTS may be ordered at: <reprints@computer.org>

Editorial production by Lisa O'Conner Cover art production by Steve Wareham Printed in the United States of America by The Printing House

IEEE Computer Society ***Conference Publishing Services*** **(CPS)**
http://www.computer.org/cps

Ontology Directed Semantic Annotation Process

Suthasinee Kuptabut
Faculty of Information Technology
King Mongkut Institute of Technology Ladkrabang
Bangkok, Thailand
suthasinee_dol@yahoo.com

Ponrudee Netisopakul
Faculty of Information Technology
King Mongkut Institute of Technology Ladkrabang
Bangkok, Thailand
ponrudee@it.kmitl.ac.th

***Abstract*—This paper proposes a process for annotating semantic concepts to sentences excerpted from webpages. The annotation process is guided by a domain specific ontology and an entities knowledge base. The overall process has four steps: extracting textual contents from a webpage, selecting relevant sentences, extracting clauses and phrases from a sentence and assigning concepts to phrases. The process is demonstrated using weather news webpages.**

Keywords-component; semantic analysis; semantic annotation; ontology; knowledge base

I. INTRODUCTION

This paper proposes a new semantic tagging process, which is a part of ontology directed semantic based parser information extraction system. The purpose of the system is to extract information from news webpages and represent semantic information in a frame-like format. In this paper, we focus on a semantic tagger. A semantic tagger assigns concepts to phrases. Concepts or semantic tags are predefined in a domain specific ontology. By referencing to classes, instances, and relations, a semantic tag helps to relate words to their relevant concepts from the domain specific ontology. The output of the semantic tagger will be further processed by a semantic analysis module. The finished system will be able to automatically construct a knowledge base from web contents to be used for other applications such as a question answering system.

Previous researches that worked on semantic tags are [1][2][3]. [1] matches semantic tags to textual contents using regular expression patterns. [2] uses a dictionary to analyze words concepts. [3] uses a domain ontology to annotate text. In our work, the semantic tagging process is guided by a domain specific ontology and a name entity extraction tool. An ontology determines the classes, their properties and relations between classes in the domain. An advantage is that using an ontology is more flexible than using a dictionary. The name entity extraction tool helps to recognize proper names, such as places and organizations in sentences. Hence, the proper names will be correctly annotated.

The paper is organized as follows. The second section reviews related research. The third section describes the overall process, the fourth section shows the experiment using a weather forecast domain and the last section is the conclusion.

II. RELATED WORK

OnTeA [1] is a semi-automatic semantic annotation tool. It is able to analyze a document such as HTML and a plain text. It mainly relies on regular expression patterns matching method. In the example, it annotates detected words using a job offer ontology. After the process, sentences will be represented as an instance of the ontology and the detected words will be the instance's properties.

[2] proposes the semantic analysis system to analyze a sentence meaning from the spoken utterances and store it into a semantic database. The semantic analysis system is developed as a part of the spoken dialog system for Croatia's weather data. The semantic analysis system analyzes sentences based on predefined semantic categories and dictionary. The semantic categories represent categories in the weather forecast domain such as the land weather forecast, the sea weather forecast and so on. The predefined dictionary is composed of vocabularies organized by semantic categories. The system assigns a semantic category to the text and extracts words using the defined dictionary. The extracted words are stored in a weather knowledge base.

[3] develops a semantic tagger that is a part of an information extraction system in the maritime search and rescue domain (SAR). The semantic tagger selects and annotates words according to the SAR ontology. If a word in the text matches an instance concept from the ontology, the semantic tagger then assigns the concept to the word. If the selected words are not found in the ontology, the semantic tagger attempts to find similar concepts to match with the selected words using the similarity measure.

Our approach is similar to previous works. However, we use name entity extraction tool to recognize proper names, such as places, organizations, and persons and regular expression patterns.

III. THE PROCESS

Fig. 1 shows an architecture of the proposed process, which consists of a webpage content extraction module, a sentence selection module, clauses and phrases extraction module and a semantic tagging module. An input to the system is a weather news webpage. The knowledge bases of the system are a weather forecast domain ontology and an entities knowledge base. The content extraction module extracts textual contents from the webpage using web content extraction tool from [6].

Then, the sentence selection module selects the related sentences by mapping words with instances from the domain ontology. Next, the clauses and phrases extraction module extracts clauses and phrases from the sentence using the Stanford parser [5]. Finally, the semantic tagging module annotates a semantic tag to a phrase.

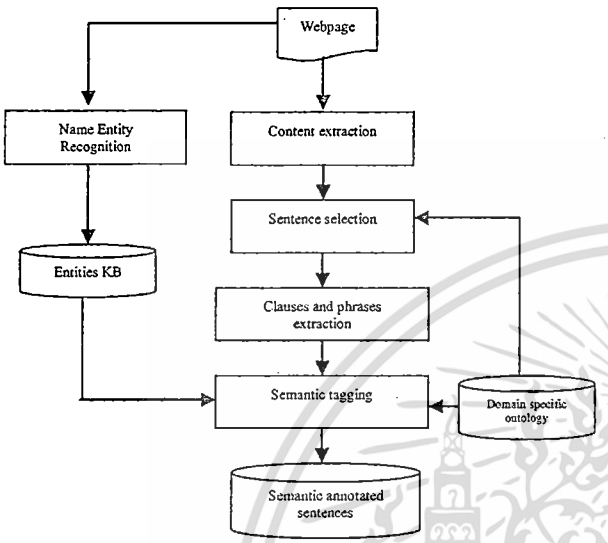


Figure 1. The architecture of a semantic tagging process

A. The domain ontology

The knowledge bases are necessary for defining semantic background of the domain. This paper uses a weather forecast domain ontology to describe related concepts in the domain.

The ontology represents concepts and properties or relationships of concepts in the domain. This paper assigns weather concepts to words in a sentence. A weather ontology is defined as weather events, which have quantifiers such as intensity, distribution, direction and so on. It also has date and time of the event and location where the event occurs. Fig. 2 illustrates the weather ontology defined by our system.

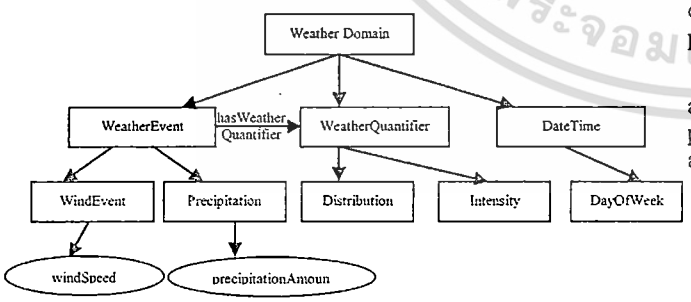


Figure 2. The example of the weather ontology

Table I shows examples of properties of class in the weather ontology.

TABLE I. THE EXAMPLE OF PROPERTIES OF CLASSES IN THE WEATHER ONTOLOGY

Concept	Property
WeatherEvent	
StormEvent	windSpeed, windMaxSpeed
PercipitationEvent	precipitaionAmount
WaveEvent	waveHeight
PressureEvent	pressureLevel
TemperatureEvent	temperatureDegrees

B. Name entity extraction

Our system uses name entity extraction tool to extract people, organizations and places. To extract these entities, we use a Stanford Named Entity Recognizer [4]. The outputs of the Stanford Named Entity Recognizer are stored in the entities knowledge base as text files format.

C. A webpage content extraction module

The input of the system is an html webpage. Only the text content in the webpage is selected. The other types of contents such as multimedia contents or graphic are left out. In addition, advertisements are removed too. The AlchemyAPI [6] provides API to extract a textual content from a webpage.

D. Sentence selection module

The textual content from the previous module may contain both relevant and irrelevant sentences to the weather domain. Therefore, the sentence selection module filters out irrelevant sentences. The sentence selection is guided by the domain specific ontology and rules. The rules define relevant concepts in the ontology for the selection process. The instances of concepts are used to classify the sentences. For example, a relevant concept is a weather concept. If the sentence contains instances of weather concepts, then the sentence is selected and sent to the next module.

E. Clauses and phrases extraction module

This module extracts clauses and phrases that correspond to concepts in the domain ontology. First, the selected sentence from the previous step is analyzed for its syntactic structure using Stanford Parser [5] with probability context free grammar (PCFG). The parser produces a parse tree. From the tree, clauses and phrases are detected. A clause is detected by pruning the parse tree. The S and SBAR tags are extracted.

The clause is then further broken into noun phrase, adjective phrase and adverb phrase. Fig. 3 illustrates the noun phrase extraction "the New England coast", "Saturday" and "wind and rain".

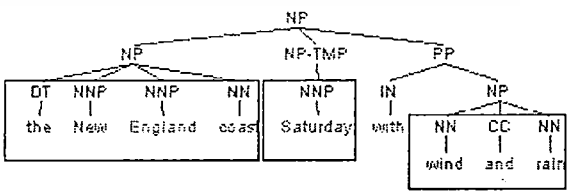


Figure 3. Excerpt of a parse tree for noun phrases extraction

Fig. 4(a) shows the phrases extracted from the sentence "Hurricane Bill spun northward toward the New England coast Saturday with wind and rain as officials warned beach lovers to head indoors for the night". There are 2 clauses: the first clause is "Hurricane Bill spun northward toward the New England coast Saturday with wind and rain", and the second clause is "officials warned beach lovers to head indoors for the night". The meaning of the first clause is related to a weather domain because it is corresponding to a hurricane weather event. Therefore, the first clause is selected. But the second clause is not selected because there is no mention of any weather event.

Fig. 4(b) shows the phrases extracted from the sentence "Forecasters said the island can expect 1 to 2 inches of rain, and Outer Cape Cod and Martha's Vineyard may receive only an inch". There are 3 clauses: the first clause is "Forecasters said", the second clause is "the island can expect 1 to 2 inches of rain", and the third clause is "Outer Cape Cod and Martha's Vineyard may receive only an inch". The second clause has a weather event of "rain", while the third clause has a weather description of "rain". Therefore, both the second clause and the third clause are extracted and sent to the next module.

F. Semantic tagging module

This module annotates extracted phrases with semantic tags using an entities knowledge base and the domain ontology. For the weather domain, the semantic tag types are weather event, weather description, weather location, weather date and weather quantifier concepts.

There are three methods for a phrase to be annotated: assigning a semantic tag by an entities knowledge base, assigning a semantic tag using regular expression patterns, and assigning a semantic tag with instances of classes in the domain ontology. Each method will be applied successively.

First, the phrase is tagged with a concept corresponding to a vocabulary in the entities knowledge base which contains the location, person and organization name. If the phrase is found in the entities knowledge base, the found phrase is tagged with its entity name. For example, the phrase "the New England coast" is annotated with the location tag.

Second, a phrase is assigned a semantic tag using regular expression patterns that define data properties of classes. For example, a phrase "1 to 2 inches" is identified by a regular expression pattern "([0-9]*[\.]*[0-9]*)*\s(inches|inch)" in which a pattern is defined in a precipitation data property of a precipitation class. Therefore, the phrase is tagged with the data property name.

Third, the phrase is searched with instances of classes in the ontology. If the phrase is found, the found phrase is assigned a semantic tag with the concept of instance.

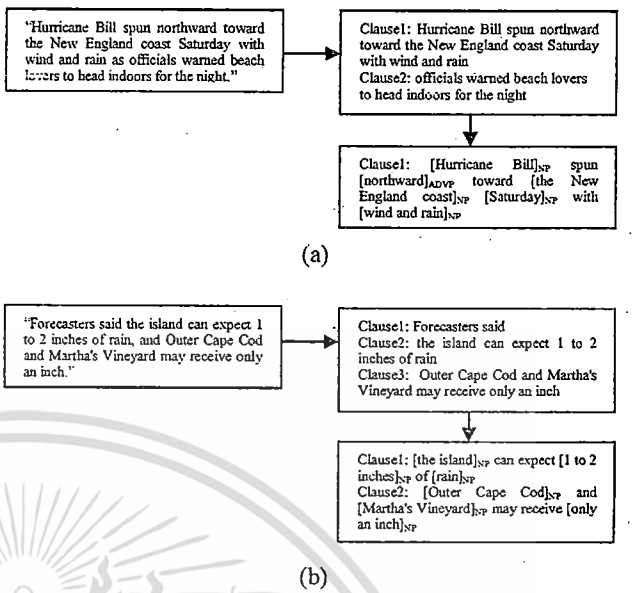


Figure 4. The example of Clauses and Phrases Extraction module (a) sentence: "Hurricane Bill spun northward toward the New England coast Saturday with wind and rain as officials warned beach lovers to head indoors for the night." (b) sentence : "Forecasters said the island can expect 1 to 2 inches of rain, and Outer Cape Cod and Martha's Vineyard may receive only an inch."

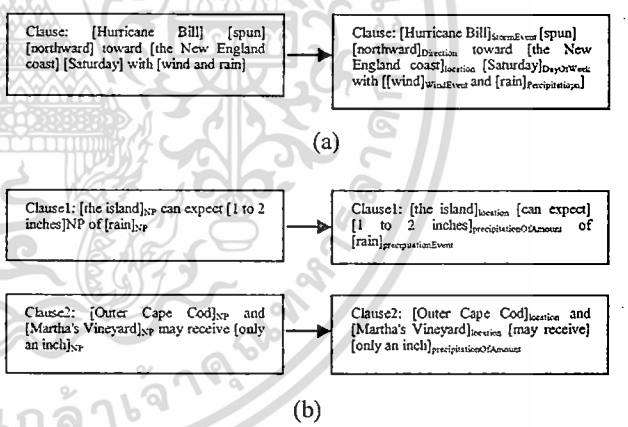


Figure 5. The example of the annotated phrases (a) sentence: "Hurricane Bill spun northward toward the New England coast Saturday with wind and rain as officials warned beach lovers to head indoors for the night." (b) sentence : "Forecasters said the island can expect 1 to 2 inches of rain, and Outer Cape Cod and Martha's Vineyard may receive only an inch."

Fig. 5(a) shows the annotated phrases in Fig. 4(a). A "Hurricane Bill", "northward", "the New England coast", "Saturday", "wind", and "rain" are annotated with a StormEvent, a Direction, a location, a DayofWeek, a WindEvent, and a Precipitation tags respectively.

Fig. 5(b) shows the annotated phrases in Figure 4(b). There are two clauses to annotate phrases. The first clause consists of phrases "the island", "1 to 2 inches", and "rain". Their phrases

are annotated with a location, a precipationAmount, and Precipitation tags respectively.

IV. Experiments and Evaluation

In our experiments, we work with weather news web pages. We design an ontology that represents concepts of a weather domain. The semantic tags are derived from the weather ontology. We collected 50 web pages containing 655 sentences.

The implement of the system is based on python programming language using RDFLib [7] and NLTK [8] library. The RDFLib is a python library for working with RDF ontology. The NLTK library works on sentence segmentation in sentence selection module and clauses and phrases extraction module.

A semantic annotation process starts working with collected webpages. These webpages are collected from yahoo weather news (http://news.yahoo.com/science/weather) and an accuweather website (http://www.accuweather.com). Textual contents of webpages are extracted using AlchemyAPI tool. The system then calls the Stanford Name Entity tool to recognize places, organizations, and persons. The results are divided into text files corresponding to various types of name entities, called an entities knowledge base. Now, the content of a webpage is broken into sentences and each sentence is sent to the sentence selection process. In our experiment, the sentence selection module selected 489 sentences out of 655 sentences for semantic annotation. Clauses and phrases that correspond to concepts in the domain ontology are extracted from the selected sentences. The number of selected clauses is 641 clauses out of 1,203 clauses. The selected clauses are then broken into phrases. Then, the phrases are annotated with semantic tags correspond to concepts in our weather domain ontology and vocabularies in the entities knowledge base previously constructed. The output of the system is semantic annotated sentences.

Table II shows the results of a sentence selection module, clauses extraction module and semantic tagging module. The second column are the total number of relevant sentences, relevant clauses and relevant phrases from webpages. The third column are numbers of selected sentences, selected clauses and selected phrases extracted by our system. The fourth column are numbers of relevant sentences, relevant clause and relevant phrases successfully extracted and annotated by our system. From the experiment, the system can extract 96.37% of relevant sentences from all relevant sentences, extract 86.63% of clauses out of relevant clauses and correctly annotate 91.22% of selected phrases and 84.63% of all relevant phrases.

TABLE II. THE RESULTS OF SENTENCE SELECTION MODULE (THE NUMBER OF SENTENCES), CLAUSES EXTRACTION MODULE (THE NUMBER OF CLAUSES) AND SEMANTIC ANNOTATION MODULE (THE NUMBER OF SEMANTIC TAGS ASSIGNED TO PHRASES)

Module	All relevant sentences from webpages	All selected sentences by our system	All selected relevant sentences from our system
Sentence selection module	496	489	478
	All relevant clauses from webpages	**All selected clauses by our system**	**All selected relevant clauses by our system**
Clauses extraction module	673	641	583
	All relevant phrases from webpages	**All selected annotated phrases by our system**	**All selected relevant annotated phrases by our system**
Semantic tagging module	2,049	1,901	1,734

To evaluate a performance, precision, recall and f-measure are calculated as follows. A precision is defined as the ratio of relevant extracted sentences, clauses and correctly annotated phrases by the system and the total number of selected sentences, clauses and phrases by the system.

$$\text{precision} = \frac{\text{relevant selections or annotation by our system}}{\text{all selections by our system}} \quad (1)$$

A recall is defined as the ratio of relevant extracted sentences, clauses and correctly annotated phrases by the system and the total number of corresponding relevant sentences, clauses or phrases from webpages.

$$\text{recall} = \frac{\text{relevant selections or annotation by our system}}{\text{all relevancies from webpages}} \quad (2)$$

To obtain a better measure to describe performance, we also used F-measure which combined precision and recall measures. The F-measure is defined as follows:

$$\text{f-measure} = \frac{2 \text{ x precision x recall}}{\text{precision} + \text{recall}} \quad (3)$$

Table III shows the precision, recall, and f-measure values for each module for ontology directed semantic annotation process. The values show that the performance of the semantic tagging module depends on the clauses extraction module. In the future, we will improve the method in clauses extraction module in order to improve the performance of the system.

A

TABLE III. THE PRECISION, RECALL, AND F-MEASURE VALUES FOR THE MODULES OF OUR SYSTEM

Module	Precision	Recall	F-measure
Sentence selection module	97.75	96.37	97.06
Clauses extraction module	90.95	86.63	88.74
Semantic tagging module	91.22	84.63	87.9

V. CONCLUSION AND FUTURE WORK

An ontology directed semantic annotation process described here is an important part of an ontology directed semantic based parser information extraction system. The process annotated phrases with semantic tags corresponding to predefined concepts in a domain specific ontology. From an experiment with weather news webpages, the proposed process is able to correctly annotate about 87.9%, in F-measure.

In future work, we will extend the methodology to cover wider domain such as natural disaster warning news and so on.

REFERENCES

[1] M. Laclavik, M. Seleng, E. Gatial, Z. Balogh, and L. Hluchy, "Ontology based Text Annotation – OnTeA," Proceeding of the 2007 conference on Information Modelling and Knowledge Bases XVIII. pp. 311-315, , 2007.

[2] A. Mestrovic, S. Martincic-Ipsic, and M. Cubrilo, "Weather Forecast Data Semantic Analysis in F-Logic," Journal of Information and Organization Sciences,Vol. 31, no. 1,pp. 115-129, 2007.

[3] N. Boufaden, "An Ontology-based Semantic Tagger for IE System," Proceeding of the 41st Annual Meeting on Association for Computational Linguistics, , Vol. 2,pp. 7-14, Sapporo, Japan, 2003.

[4] Stanford Name Entity, http://nlp.stanford.edu/software/CRF-NER.shtml

[5] Stanford Parser, http://nlp.stanford.edu/software/lex-parser.shtml

[6] AlchemyAPI , http://www.alchemyapi.com/

[7] RDFLib, http://www.rdflib.net/

[8] NLTK (Natural Language Toolkit), http://www.nltk.org/