วิทยานิพนธ์

การพัฒนาระบบสืบค้นข้อมูลบุคคลจากเวิลด์ ไวด์ เว็บ ด้วยเทคนิคอนุมาน

A Development of Person Information Search from WWW based on Inference Technique

นายดุษฎี ธรรมวิจิตร

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

พ.ศ. ๒๕๔๘

ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

ปริญญา

วิศวกรรมคอมพิวเตอร์	วิศวกรรมคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง การพัฒนาระบบสืบค้นข้อมูลบุคคลจาก เวิลด์ ไวด์ เว็บ ด้วยเทคนิคอนุมาน

A Development of Person Information Search from WWW based on Inference Technique

นามผู้วิจัย นายดุษฎี ธรรมวิจิตร

ได้พิจารณาเห็นชอบโดย

ประธานกรรมการ .....

(รองศาสตราจารย์อัศนีย์ ก่อตระกูล, D.Eng.)

กรรมการ .....

(อาจารย์จิตร์ทัศน์ ฝักเจริญผล, Ph.D.)

กรรมการ .....

(อาจารย์ยอดเยี่ยม ทิพย์สุวรรณ์, Ph.D.)

หัวหน้าภาควิชา .....

(อาจารย์พีรวัฒน์ วัฒนพงศ์, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์วินัย อาจคงหาญ, M.A.)

คณบดีบัณฑิตวิทยาลัย

วันที่ 5 **เดือน** เมษายน **พ.ศ.** 2549

วิทยานิพนธ์

เรื่อง

การพัฒนาระบบสืบค้นข้อมูลบุคคลจากเวิลด์ ไวด์ เว็บ ด้วยเทคนิคอนุมาน

A Development of Person Information Search from WWW based on Inference Technique

โดย

นายดุษฎี ธรรมวิจิตร

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
พ.ศ. 2549

ISBN 974-16-1441-1

ดุษฎี ธรรมวิจิตร 2549: การพัฒนาระบบสืบค้นข้อมูลบุคคลจาก เวิลด์ ไวด์ เว็บ ด้วยเทคนิคอนุมาน ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ ประธานกรรมการที่ปรึกษา: รองศาสตราจารย์อัศนีย์ ก่อตระกูล, D.Eng. 70 หน้า
ISBN 974-16-1441-1

ในปัจจุบันข้อมูลในระบบอินเทอร์เน็ตหรือเวิลด์ ไวด์ เว็บ นั้นได้เพิ่มจำนวนขึ้นเป็นจำนวนมาก ด้วยความรวดเร็ว รวมถึงข้อมูลบุคคลซึ่งเป็นสิ่งที่ผู้ใช้ต้องการค้นหามากถึง 30 % วิธีการที่เป็นที่นิยมสำหรับใช้เพื่อค้นหาข้อมูลบุคคลจากเวิลด์ ไวด์ เว็บ คือ การค้นหาข้อมูลผ่านบริการของเสิร์ชเอนจิ้น อย่างไรก็ตาม วิธีค้นหาข้อมูลในลักษณะนี้ ทำให้ผู้ใช้ต้องเสียเวลานานในการค้นหาข้อมูล โดยเวลาที่สูญเสียไปเกิดขึ้นจาก 2 สาเหตุคือ เวลาที่ใช้ในการเลือกคำค้นที่เหมาะสมเพื่อให้ได้เอกสารที่มีข้อมูลของบุคคลที่ต้องการ และเวลาที่ต้องใช้ในการตรวจสอบผลลัพธ์จากการค้นคืนของเสิร์ชเอนจิ้นว่าเป็นข้อมูลที่ผู้ใช้ต้องการหรือไม่ อีกทั้งข้อมูลบุคคลที่ต้องการอาจจะกระจัดกระจายเว็บเพจต่างๆทำให้เสียเวลาในการรวบรวมอีกด้วย ดังนั้นงานวิจัยนี้จึงวิจัยและพัฒนาระบบสืบค้นข้อมูลบุคคลจากเวิลด์ ไวด์ เว็บ โดยระบบนี้จะรวบรวมเว็บเพจที่มีข้อมูลบุคคลปรากฏอยู่แล้วสกัดเฉพาะข้อมูลส่วนที่เกี่ยวกับบุคคลเท่านั้น เช่น ชื่อ นามสกุล ตำแหน่ง องค์กรที่สังกัด โดยก่อนจัดเก็บข้อมูลที่สกัดได้จะมีการตรวจสอบชื่อของบุคคลว่าชื่อใดเป็นชื่อของบุคคลคนเดียวกันบ้าง เพื่อที่จะรวมข้อมูลเข้าด้วยกัน และเปิดบริการให้ผู้ใช้สามารถสืบค้นข้อมูลเหล่านี้ได้ทั้งทางตรงและทางอ้อม โดยสำหรับการเข้าถึงข้อมูลโดยทางอ้อมนั้นจะใช้เทคนิคอนุมานแบบอาศัยกฎ ซึ่งจะกระทำบนข้อมูลที่เก็บอยู่ในโครงสร้างแบบอาร์ดีเอฟ (RDF) เพื่อช่วยให้ผู้ใช้สามารถเข้าถึงข้อมูลเหล่านี้ได้

สิ่งที่พัฒนาขึ้นในระบบสืบค้นข้อมูลบุคคลในงานวิจัยนี้มี 3 ส่วนได้แก่ ส่วนรวบรวมเว็บเพจที่มีข้อมูลบุคคล ส่วนรวมข้อมูลที่สกัดมาได้เข้าไว้ด้วยกัน โดยจะตรวจสอบหาชื่อคนที่คล้ายกันว่าเป็นชื่อของคนๆเดียวกันหรือไม่ ซึ่งได้พัฒนาจากวิธีการตรวจสอบความคล้ายของสตริง (Edit distance) เพิ่มเติมให้เหมาะสมกับภาษาไทย และส่วนของการให้บริการสืบค้นข้อมูล ซึ่งในส่วนนี้จะมีการเพิ่มเทคนิคอนุมานแบบอาศัยกฎ และออนโทโลจี (Ontology) ซึ่งใช้ในการจัดเก็บความรู้เกี่ยวกับตำแหน่งของบุคคล และสาขาความเชี่ยวชาญในหน่วยงานมหาวิทยาลัย เพื่อช่วยให้ผู้ใช้สามารถเข้าถึงข้อมูลที่เกิดจากการสรุปได้

เมื่อทดลองกับเว็บเพจที่รวมมาได้ทั้งสิ้นจำนวน 169,079 เว็บเพจ ผลปรากฏว่า ส่วนรวมข้อมูลบุคคลที่สกัดได้เข้าไว้ด้วยกันให้ค่าความเที่ยง 0.85 และค่าระลึกได้ 0.88 เมื่อใช้ค่าขอบเขตเท่ากับ 1.5 สำหรับส่วนของการสืบค้นข้อมูลให้ค่าความถูกต้อง 0.83 และค่าระลึกได้ 0.85 เมื่ออาศัยออนโทโลจีและกฎการอนุมานแล้วค่าความถูกต้องเพิ่มขึ้น 0.04 ค่าระลึกได้เพิ่มขึ้น 0.22

ดุษฎี ธรรมวิจิตร — ลายมือชื่อนิสิต

ลายมือชื่อประธานกรรมการ — 19 / 03 / 49

Dussadee Thamvijit 2006: A Development of Person Information Search from WWW based on Inference Technique. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Asanee Kawtrakul, D.Eng. 70 pages.
ISBN 974-16-1441-1

Nowadays, the amount of useful information provided by the Internet is currently growing at an incredible rate. Recently, searching for person information from the Internet is 30% of all search queries. The method usually used for searching person information from WWW is to use traditional search engine. However, users must spend a lot of time finding appropriate query words, reading retrieved documents to find required information, and, in some situations, gathering information from various sources. Therefore, this research proposed a person information searching system from WWW that automatically collects web pages having person information, and extracts person information (e.g. name, surname, position, and organization) from collected web pages. The system also provides a service for searching person information both direct and indirect way. To search for the information indirectly, the system use a rule-based reasoning technique that processes information stored in RDF model.

Three important components of person information searching system consisting of person information gathering, person information integration, and information accessing are developed in this research. A new edit distance based similarity checking method is proposed to check the similarity of Thai person name in information integration module. The information accessing component uses a reasoning technique to access required information indirectly.

The experiment with 169,079 collected web pages from university and organization shows that the precision and recall of the information integration module are 0.85 and 0.88, and the precision and recall of the information accessing module are 0.83 and 0.85 respectively. Moreover, when using ontology and reasoning technique the precision and recall are increased by 0.04 and 0.22.

Dussadee Thamvijit
Student's signature

Asanee Kawtrakul
Thesis Advisor's signature

29 / 03 / 06

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลงได้ด้วยความช่วยเหลือจากบุคคลหลายท่าน ข้าพเจ้าขอขอบพระคุณ รองศาสตราจารย์ ดร. อัศนีย์ ก่อตระกูล ประธานกรรมการที่ปรึกษา ผู้ให้แนวทางการวิจัย ให้ข้อเสนอแนะที่เป็นประโยชน์ อาจารย์ ดร. จิตร์ทัศน์ ฝักเจริญผล กรรมการที่ปรึกษาวิชาเอก อาจารย์ ดร. ยอดเยี่ยม ทิพย์สุวรรณ์ กรรมการที่ปรึกษาวิชารอง และอาจารย์ณัฐวุฒิ ขวัญแก้ว อาจารย์ผู้แทนบัณฑิตวิทยาลัย ที่กรุณาให้คำปรึกษา และข้อเสนอแนะที่มีคุณค่า เพื่อให้วิทยานิพนธ์นี้สมบูรณ์ยิ่งขึ้น

ข้าพเจ้าขอขอบคุณ เพื่อนๆ และพี่ๆ ห้องปฏิบัติการวิจัยเชี่ยวชาญเฉพาะการประมวลผลภาษาธรรมชาติและเทคโนโลยีสารสนเทศอัจฉริยะ (NAiST Lab) ทุกท่านที่ให้คำปรึกษาเกี่ยวกับการทำงานวิจัย และเจ้าหน้าที่ธุรการภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ทุกท่านสำหรับความช่วยเหลือด้านการประสานงาน และดำเนินการเอกสารต่างๆ

ดุษฎี ธรรมวิจิตร

เมษายน 2549

สารบัญ

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ (ต่อ)

การพัฒนาระบบสืบค้นข้อมูลบุคคลจากเวิลด์ ไวด์ เว็บ ด้วยเทคนิคอนุมาน

A Development of Person Information Search from WWW based on Inference Technique

คำนำ

เนื่องจากในปัจจุบันอินเทอร์เน็ตเป็นแหล่งข้อมูลที่มีขนาดใหญ่และมีข้อมูลหลากหลายประเภท วิธีที่เป็นที่นิยมสำหรับการสืบค้นข้อมูลบนอินเทอร์เน็ตคือ การสืบค้นผ่านบริการเสิร์ชเอนจิ้น ซึ่งหัวข้อหนึ่งที่น่าสนใจสำหรับการสืบค้นคือ การค้นหาข้อมูลบุคคล ตัวอย่างเช่น “ต้องการข้อมูลสำหรับติดต่อนายสมชาย”, “ใครเป็นผู้เชี่ยวชาญด้านการประมวล ผลภาษาด้วยคอมพิวเตอร์” และ “น.ส.สมหญิงทำงานอะไร” เป็นต้น ซึ่งการหาข้อมูลเพื่อตอบคำถามข้างต้นนั้นเป็นงานที่ต้องใช้เวลาเป็นอย่างมาก ทั้งนี้เนื่องจากเสิร์ชเอนจิ้นที่มีอยู่ในปัจจุบันไม่ได้ออกแบบมาเพื่อตอบคำถามดังกล่าวโดยตรง โปรแกรมจะทำหน้าที่ค้นหาเอกสารที่มีคำค้นของผู้ใช้เท่านั้น ทำให้ผู้ใช้ต้องเสียเวลาในการอ่านเอกสารเพื่อค้นหาข้อมูลที่ต้องการด้วยตนเอง ซึ่งข้อมูลที่ต้องการนั้นอาจกระจัดกระจายอยู่ตามเอกสารต่างๆ ทำให้ต้องเสียเวลาในการรวบรวมข้อมูลจากเอกสารจำนวนมาก จึงจะได้ข้อมูลครบถ้วนตามที่ต้องการ ยิ่งไปกว่านั้น ในกรณีที่ผู้ใช้ไม่ใช่ผู้เชี่ยวชาญในหัวข้อที่ต้องการสืบค้น เอกสารที่ได้อาจไม่ตรงตามความต้องการ เนื่องจากผู้ใช้ใช้คำค้นที่ไม่ถูกต้อง ตัวอย่างเช่น ผู้ใช้ต้องการค้นหาผู้เชี่ยวชาญด้านคอมพิวเตอร์ ซึ่งปัญหา คือ ต้องใช้คำค้นใดจึงสามารถค้นหาระดับของความเชี่ยวชาญของผู้เชี่ยวชาญด้านคอมพิวเตอร์ได้ เป็นต้น

ดังนั้นปัญหาที่เกิดขึ้นจากการสืบค้นข้อมูลบุคคลได้แก่ ปัญหาการได้มาของข้อมูลบุคคลไม่ครบถ้วนตามจำนวนที่มีอยู่จริง อันมีสาเหตุเนื่องมาจากการกระจายตัวของข้อมูลบุคคล และปัญหาการเลือกใช้คำค้นของผู้ใช้ซึ่งไม่ตรงกันกับข้อมูลที่มีอยู่

ปัญหาการได้มาของข้อมูลบุคคลไม่ครบถ้วนตามจำนวนที่มีอยู่จริงนั้น เกิดขึ้นจากลักษณะของข้อมูลที่อยู่ในแต่ละเว็บเพจมีการนำเสนอข้อมูลของแต่ละบุคคลในลักษณะแตกต่างกัน เช่น การนำเสนอข้อมูลในรูปแบบตารางหรือลิสต์ ซึ่งถ้าเราต้องการให้คอมพิวเตอร์สามารถเข้าใจข้อมูลที่อยู่ในเว็บเพจได้นั้น จำเป็นต้องมีการเปลี่ยนแปลงโครงสร้างของเว็บเพจให้มีลักษณะโครงสร้างที่

คอมพิวเตอร์สามารถเข้าใจได้ โดยระบบได้แบ่งลักษณะโครงสร้างของเว็บเพจบุคคลออกเป็น 3 โครงสร้าง คือ ตาราง, ลิสต์ และข้อความ ซึ่งโครงสร้างแต่ละแบบจะใช้วิธีในการสกัดข้อมูลที่แตกต่างกันไป หลังจากนั้นระบบจึงนำผลลัพธ์ที่ได้จากการสกัดข้อมูลมารวมกัน นอกจากนี้ในส่วนของการรวบรวมข้อมูลของคนๆ เดียวกันเข้าด้วยกัน ยังเกิดปัญหา การสะกดชื่อผิด เช่น คำว่า “ทักษิณ ชินวัตร” สะกดเป็น “ทักษิน ชินวัตร” โดยมีการใช้ตัวสะกด “ณ” เป็น “น” และปัญหาการย่อชื่อที่มีความยาวมากให้สั้นลง เช่น “มาริสา สุโกศล หนุนภักดี” เขียนย่อเป็น “มาริสา สุโกศล” ซึ่งเป็นปัญหาที่ทำให้ข้อมูลของบุคคลๆ เดียวกัน ไม่สามารถรวมเข้าไว้เป็นหน่วยเดียวกันได้ จึงส่งผลให้ข้อมูลมีความซ้ำซ้อนกัน จากการสำรวจพบว่าปัญหาการสะกดชื่อผิดเป็นปัญหาที่พบมาก ดังนั้นระบบนี้จึงมุ่งเน้น การแก้ปัญหาชื่อที่สะกดผิดเพียงอย่างเดียว โดยใช้วิธีการตรวจสอบความคล้ายที่รองรับการทำงานกับภาษาไทย ซึ่งเป็นวิธีที่นำวิธีการวัดความแตกต่างของสตริง (Edit distance) มาประยุกต์ใช้โดยเพิ่มกฎและปรับปรุง การทำงานให้สอดคล้องกับลักษณะของภาษาไทย

ปัญหาการเลือกใช้คำค้นของผู้ใช้ซึ่งไม่ตรงกันกับข้อมูลที่มีอยู่ เกิดขึ้นจากผู้ใช้บางคนอาจไม่ได้มีความเชี่ยวชาญหรือความคุ้นเคยในหัวข้อที่ต้องการค้นหาอยู่ เป็นผลทำให้ผู้ใช้อาจเลือกคำค้นไม่ตรงกับคำในเอกสารที่มีอยู่จริง ในขณะที่เอกสารที่มีอยู่จริงนั้นมีข้อมูลที่ผู้ใช้ต้องการอยู่

จากปัญหาการค้นหาข้อมูลบุคคลโดยผ่านบริการเสิร์ชเอนจิ้นที่กล่าวมาข้างต้น งานวิจัยนี้จึงได้ศึกษาปัญหาและพัฒนาระบบสำหรับสกัดและให้บริการในการสืบค้นสำหรับข้อมูลบุคคล ซึ่งระบบนี้ประกอบด้วยส่วนประกอบ 4 ส่วนด้วยกันคือ 1) ส่วนรวบรวมข้อมูล ซึ่งทำหน้าที่ในการรวบรวมเว็บเพจที่มีข้อมูลของบุคคลจากหลายแหล่งข้อมูลมารวมไว้ที่เดียวกัน 2) ส่วนสกัดข้อมูล ทำหน้าที่แปลงเอกสารเว็บเพจที่อยู่ในรูปแบบเอชทีเอ็มแอล ให้อยู่ในโครงสร้างที่เหมาะสมต่อการประมวลผลด้วยคอมพิวเตอร์ และยังทำหน้าที่ในการสกัดข้อมูล 3) ส่วนรวมข้อมูลของคนๆเดียวกันเข้าด้วยกัน เป็นการกรองข้อมูลของบุคคลหลังจากที่สกัดออกมากได้ในแต่ละหน่วย เพื่อตรวจสอบดูว่าข้อมูลของคนๆเดียวกันนั้นได้ถูกรวมอยู่เป็นหน่วยเดียวกันหรือไม่ 4) ส่วนสืบค้นโดยอาศัยการอนุมานและออนโทโลจี ซึ่งจะช่วยเพิ่มช่องทางในการค้นหาข้อมูลให้มากขึ้น โดยในส่วนนี้แม้ว่าผู้ใช้เลือกใช้คำค้นไม่ตรงกันกับคำที่มีอยู่ในข้อมูลที่ถูกจัดเก็บไว้แล้ว ถ้าคำค้นนั้นมีความสัมพันธ์กับข้อมูลที่ถูกจัดเก็บแล้ว ข้อมูลเหล่านั้นจะถูกค้นคืนให้แก่ผู้ใช้

วัตถุประสงค์และขอบเขต

วัตถุประสงค์

1. เพื่อศึกษาเทคนิคการสืบค้นข้อมูลบุคคลที่มีประสิทธิผลทางทั้งทางด้านความถูกต้องและความครอบคลุมในการสืบค้นข้อมูลบุคคล
2. เพื่อศึกษาวิธีในการนำเทคนิคอนุมานมาประยุกต์ใช้ในการสืบค้นข้อมูล เพื่อประโยชน์ในการเข้าถึงข้อมูลโดยอ้อมได้
3. เพื่อศึกษาการนำออนโทโลจีมาใช้เพิ่มประสิทธิภาพกับระบบสืบค้นข้อมูลบุคคล
4. เพื่อศึกษาวิธีการรวบรวมข้อมูลของบุคคลคนเดียวกันเข้าด้วยกัน

ขอบเขตของงานวิจัย

1. ใช้ค้นหาบุคคลในกลุ่มนักวิชาการในหน่วยงานมหาวิทยาลัย
2. ใช้กับเอกสารภาษาไทย
3. ออนโทโลจีที่ใช้เกี่ยวกับสาขาความเชี่ยวชาญในหน่วยงานมหาวิทยาลัย
4. สามารถค้นหาบุคคลด้วยการระบุความเชี่ยวชาญ

การตรวจเอกสาร

วิธีที่ใช้ในการค้นหาข้อมูลของบุคคลโดยใช้คอมพิวเตอร์ในการสืบค้นโดยคอมพิวเตอร์นั้นสามารถกระทำได้โดยวิธีการดังนี้คือ 1) สืบค้นโดยผ่านบริการเสิร์ชเอนจิ้น ยกตัวอย่างเช่น กูเกิ้ล (http://www.google.com) หรือ ยาฮู (http://www.yahoo.com) ซึ่งในสืบค้นลักษณะนี้นั้นมีหลักการคือ ผู้ใช้จะเป็นผู้ส่งคำค้นให้กับระบบเสิร์ชเอนจิ้น หลังจากนั้นระบบจะค้นคืนเอกสารที่มีคำค้นที่ผู้ใช้เลือกกลับคืนให้แก่ผู้ใช้ ซึ่งวิธีการสืบค้นข้อมูลของบุคคลในลักษณะนี้ทำให้ผู้ใช้ต้องใช้เวลาในการตรวจสอบผลลัพธ์ที่ได้มาว่าใช่ข้อมูลของบุคคลที่ต้องการหรือไม่ และถ้าหากผลลัพธ์ที่ส่งมานั้นมีจำนวนมาก ทำให้ผู้ใช้ต้องเสียเวลาในการสืบค้นมากตามไปด้วย ดังนั้นงานวิจัยนี้จึงนำเสนอระบบที่สามารถรวบรวมข้อมูลของบุคคลที่มีอยู่ตามเว็บเพจต่างๆ มาไว้ที่เดียวกันและเปิดให้บริการในการสืบค้นทำให้ผู้ใช้เข้าถึงข้อมูลที่ต้องการได้รวดเร็วกว่า โดยความรู้พื้นฐานที่มีความเกี่ยวข้องกับระบบสืบค้นข้อมูลบุคคลได้แก่ การเสาะหาข้อมูลที่มีอยู่ในระบบอินเทอร์เน็ต วิธีที่ใช้ในการตรวจสอบความคล้ายของสตริง และวิธีที่ใช้ในการจัดเก็บข้อมูลที่สามารถรองรับการอนุมาน

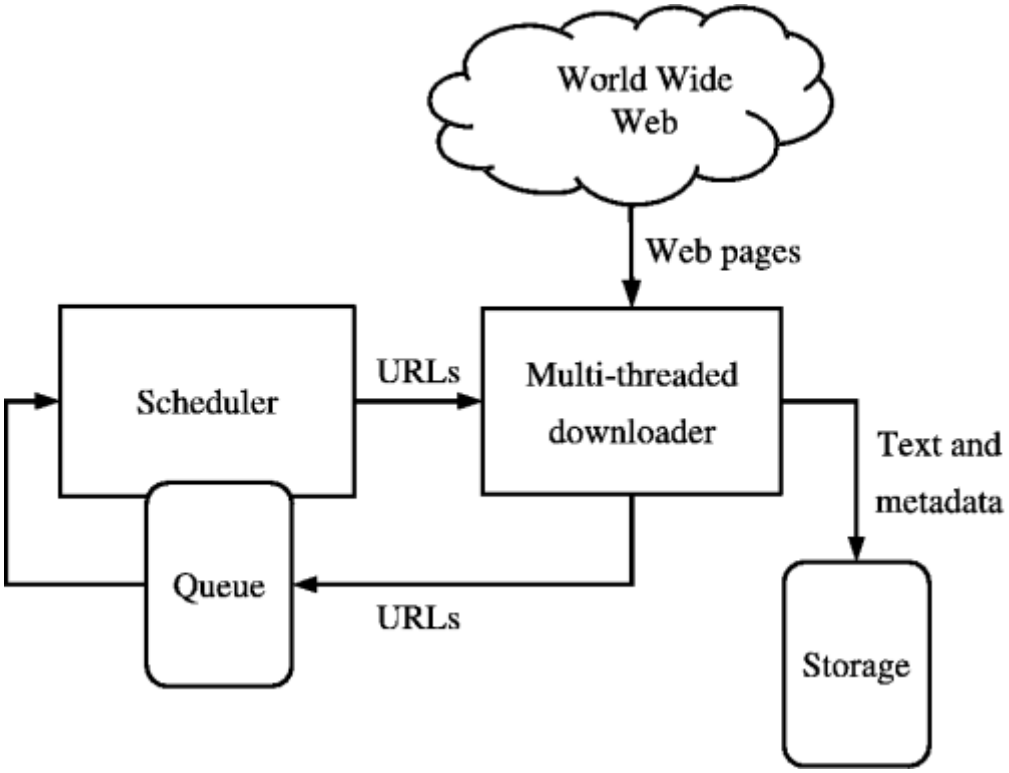
ความรู้พื้นฐาน

การเสาะหาข้อมูลที่มีอยู่ในระบบอินเทอร์เน็ต

วิธีการที่เป็นที่นิยมในการรวบรวมข้อมูลในระบบอินเทอร์เน็ตนั้นคือ เว็บครอว์เลอร์ (Web crawler) ซึ่งเป็นโปรแกรมที่รวบรวมเว็บเพจอย่างอัตโนมัติ โดยส่วนใหญ่แล้วเว็บคลอว์เลอร์จะถูกเรียกใช้งานโดยเสิร์ชเอนจิ้น (โปรแกรมใช้สำหรับให้บริการในการสืบค้นเอกสารที่มีอยู่อินเทอร์เน็ต) หน้าที่ของเว็บคลอว์เลอร์ที่มีต่อเสิร์ชเอนจิ้นคือ การบำรุงรักษาข้อมูลภายในเสิร์ชเอนจิ้นโดยการตรวจสอบความทันสมัยของข้อมูลในระบบ อีกทั้งยังช่วยนำเข้าเว็บเพจที่ยังไม่มีในระบบเข้ามาอีกด้วย

การทำงานของเว็บคลอว์เลอร์นั้นเริ่มต้นจากการอ่านลิสต์ของยูอาร์แอล (URL) เพื่อใช้เป็นจุดเริ่มต้นในการดาวน์โหลด หลังจากที่เว็บคลอว์เลอร์ได้ดาวน์โหลดเว็บเพจมาแล้ว ส่วนที่เป็นไฮเปอร์ลิงค์ (hyperlink) ทั้งหมดจะถูกสกัดออกจากเว็บเพจเหล่านั้น และไฮเปอร์ลิงค์ที่สกัดออกมาได้จะถูกอ่านโดยเว็บครอว์เลอร์เพื่อใช้เป็นยูอาร์แอลที่จะทำการดาวน์โหลดต่อไป โดยกระบวนการ

ทำงานของเว็บครอว์เลอร์ก็จะทำซ้ำวนรอบเช่นนี้เรื่อยไป ภาพจำลองสถาปัตยกรรมของเว็บครอว์เลอร์แบบทั่วไปนั้นแสดงดังภาพที่ 1



ภาพที่ 1 สถาปัตยกรรมแบบทั่วไปของเว็บครอว์เลอร์

จากภาพที่ 1 เป็นสถาปัตยกรรมของเว็บครอว์เลอร์ ซึ่งมีขั้นตอนในการทำงานดังที่ได้กล่าวมาแล้วข้างต้น

ตัวอย่างของเว็บครอว์เลอร์

ต่อไปนี้เป็นตัวอย่างของเว็บครอว์เลอร์ที่มีการเปิดเผยโครงสร้างของสถาปัตยกรรมเว็บครอว์เลอร์แต่ละตัวนั้นจะมีรายละเอียดของการทำงานแตกต่างกันไป

RBSE (Eichmann, 1994) นำเสนอเว็บครอว์เลอร์ที่มีพื้นฐานอยู่บนโปรแกรม 2 โปรแกรม โดยโปรแกรมแรกคือ สไปเดอร์ (Spider) ซึ่งมีหน้าที่รับผิดชอบเรื่องการจัดการคิวของยูอาร์แอลในฐานข้อมูล และโปรแกรมที่สองคือ ไมค์ (mite) ซึ่งเป็นโปรแกรมที่ใช้ดาวน์โหลดเว็บเพจต่างๆ

WebCrawler (Pinkerton, 1994) เป็นเว็บครอว์เลอร์ตัวแรกที่ถูกใช้งานร่วมกันกับเสิร์ชเอนจิ้นทำอินเด็กซ์กับเอกสารทั้งเอกสาร อัลกอริทึมที่ใช้ในการสำรวจเว็บเพจที่มีอยู่ในอินเทอร์เน็ตจะเป็นแบบ การค้นหาแนวกว้าง (Breath first exploration)

World Wide Web Worm (McBryan, 1994) เป็นเว็บครอว์เลอร์ที่มีการทำงานแบบง่าย ใช้เพื่อทำอินเด็กซ์ของหัวข้อของเอกสาร และยูอาร์แอล โดยอินเด็กซ์จะถูกค้นหาโดยใช้คำสั่ง “grep” ซึ่งเป็นคำสั่งของระบบปฏิบัติการแบบยูนิคส์

Google Crawler (Brin and Page, 1998) ได้อธิบายรายละเอียดบางประการของเว็บครอว์เลอร์ที่ถูกใช้เพื่อทำงานร่วมกับกูเกิ้ลเสิร์ชเอนจิ้น เว็บครอว์เลอร์นี้ถูกพัฒนาขึ้นโดยใช้ภาษาซีพลัสพลัส (C++) และ ไพธอน (Python) ขั้นตอนของการทำงานจะถูกรวมอยู่กับขั้นตอนการทำอินเด็กซ์เนื่องจากกระบวนการในการสกัดเอาไฮเปอร์ลิงค์ออกมานั้น เนื่องจากกูเกิ้ลเป็นเสิร์ชเอนจิ้นที่ต้องการเก็บข้อมูลที่มีปริมาณมาก ดังนั้นเว็บครอว์เลอร์จึงต้องรับงานในปริมาณมากตามไปด้วย ดังนั้นจึงมียูอาร์แอลเซิร์ฟเวอร์ทำหน้าที่ส่งลิสต์ของยูอาร์แอลไปให้ส่วนที่ทำหน้าที่ดาวน์โหลดเว็บเพจ ซึ่งลักษณะเด่นของส่วนนี้คือมีการตรวจสอบว่ายูอาร์แอลที่สกัดออกมาได้นั้น ซ้ำกับยูอาร์แอลที่ได้ดาวน์โหลดมาแล้วหรือไม่

CobWeb (da Silva et al 1999) ใช้เทคนิคในการกระจายตัวดาวน์โหลดให้ไปทำงานหลายตัวพร้อมๆกัน โดยมีโปรแกรมที่ช่วยในการจัดลำดับการทำงาน (Scheduler) ของโปรแกรมดาวน์โหลด (collector) ให้สอดคล้องกัน การทำงานของโปรแกรมจัดลำดับการทำงานนั้นได้ใช้วิธีการค้นหาแนวกว้าง (breadth-first search) ร่วมกันกฎระเบียบของการค้นหาแบบสุภาพ (politeness policy) เพื่อไม่ให้เซิร์ฟเวอร์ที่ถูกค้นหามากจนเกินไป โดยตัวเว็บครอว์เลอร์นี้เขียนด้วยภาษาเพิร์ล (perl)

Mercator (Heydon and Najork, 1999) เป็นเว็บครอว์เลอร์ที่เขียนด้วยภาษาจาวา ซึ่งมีการแบ่งส่วนของการทำงานออกเป็น 2 ส่วน ได้แก่ โมดูลโปรโตคอล (Protocal modules) และโมดูลการประมวลผล (Process modules) ซึ่งโมดูลโปรโตคอลมีหน้าที่เกี่ยวกับการจัดหารูปแบบของการได้มาของเว็บเพจ ตัวอย่างเช่น โดยการใช้ โปรโตคอลเอชทีทีพี ส่วนโมดูลการประมวลผลมีการทำงานเกี่ยวข้องกับการประมวลผลเว็บเพจที่หามาได้ ตัวอย่างเช่น กรองภาษาเอชทีเอ็มแอลออกจากเว็บเพจ และการสกัดไฮเปอร์ลิงค์ออกจากเว็บเพจ

WebFountain (Edwards et al, 2001) เว็บครอว์เลอร์นี้มีลักษณะคล้ายคลึงกันกับ Mercator ที่ได้กล่าวมาแล้วก่อนหน้านี้ แต่เว็บครอว์เลอร์ตัวนี้ถูกพัฒนาขึ้นโดยใช้ภาษา C++ โดยมีฟังก์ชั่นที่มีลักษณะเด่นมีดังนี้ คอลโทรลเลอร์ จะมีหน้าที่คำนวณอัตราการเปลี่ยนแปลงของเว็บเพจแต่ละเว็บเพจเพื่อตรวจสอบว่าเว็บเพจใดความถูกดาวน์โหลดเพื่อทำให้ข้อมูลมีความทันสมัย นอกจากจะทำหน้าที่ในการดาวน์โหลดเว็บเพจเข้ามาเก็บไว้ในระบบเพียงอย่างเดียวซึ่งวิธีที่ใช้ในการคำนวณนั้นใช้วิธีการแบบ non-linear programming เพื่อคำนวณว่าเว็บเพจใดที่ควรจะต้องถูกดาวน์โหลดมาใหม่ เพื่อทำให้ข้อมูลที่ได้จัดเก็บไว้นั้นมีความทันสมัยอยู่เสมอ

PolyBot (Shkapenyuk and Suel, 2002) คือเว็บครอว์เลอร์ที่มีการกระจายการทำงานพร้อมกันหลายการประมวลผล (distributed crawler) ซึ่งถูกเขียนด้วยภาษาซีพลัสพลัสและภาษาไพทอน ซึ่งประกอบด้วยฟังก์ชั่นดังต่อไปนี้คือ “crawl manager”, “downloader”, “DNS revolvers” กระบวนการทำงานมีดังนี้ ยูอาร์แอลที่ได้มานั้นจะถูกจัดเข้าคิวไว้เพื่อรอการดาวน์โหลด หลังจากนั้นจะมีโปรแกรมมาคอยตรวจสอบดูว่ายูอาร์แอลใดได้รับการดาวน์โหลดไปบ้างแล้วเพื่อจะได้ไม่ต้องดาวน์โหลดซ้ำ

WebRace (Zeinalipour-Yazti and Dikaiakos, 2002) เป็นเว็บครอว์เลอร์ที่พัฒนาขึ้นมาโดยใช้ภาษาจาวา (Java) และถูกใช้เพื่อเป็นส่วนการทำงานส่วนหนึ่งของระบบ “eRACE” ซึ่งเป็นระบบที่คอยรับคำสั่งจากผู้ใช้ว่าต้องการที่จะดาวน์โหลดเว็บเพจใด และเว็บเพจนั้นจะถูกตรวจสอบอยู่ตลอดเวลา เมื่อเว็บเพจนั้นมีการเปลี่ยนแปลงเกิดขึ้น เว็บเพจนั้นก็จะถูกดาวน์โหลดมาเก็บไว้ในระบบ

Ubicrawler (Boldi et al, 2004) เป็นเว็บครอว์เลอร์ที่พัฒนาขึ้นด้วยภาษาจาวาที่มีการกระจายการทำงานเป็นแบบหลายตัวพร้อมๆกัน แต่ไม่มีการประมวลผลกลางที่คอยควบคุมการทำงาน ข้อดีของเว็บครอว์เลอร์นี้คือ จะไม่มีการดาวน์โหลดเว็บเพจที่ซ้ำกันมา โดยหน้าที่ในการตรวจสอบว่าเว็บเพจใดซ้ำหรือไม่ซ้ำกับของเดิมนั้นจะเป็นหน้าที่ของตัวแทน (agent) อีกทั้งเว็บครอว์เลอร์นี้ยังได้ถูกออกแบบมาเพื่อให้รองรับการทำงานกับข้อมูลที่มีปริมาณมาก และมีเสถียรภาพในการทำงานสามารถรอบรับความผิดพลาดหลายอย่างที่จะเกิดขึ้นได้อีกด้วย

FAST Crawler (Risvik and michelsen, 2002) เป็นเว็บครอว์เลอร์ของระบบเสิร์ชเอนจิ้นที่มีชื่อว่า FAST เสิร์ชเอนจิ้นและขั้นตอนการทำงานของเว็บครอว์เลอร์นี้ได้รับการเปิดเผยสู่สาธารณะ

ส่วนที่ใช้ในการจัดการคิวของยูอาร์แอลที่จะต้องดาวน์โหลดนั้นมีชื่อว่า “document processor” เว็บครอว์เลอร์แต่ละตัวที่ถูกกระจายการทำงานออกไปนั้นได้ติดต่อสื่อสารกันผ่านโมดูลกระจายการทำงาน (distributor module) โดยการแลกเปลี่ยนข้อมูลของไฮเปอร์ลิงค์ที่มีอยู่จะถูกกระทำโดยโมดูลนี้

จากลักษณะของเว็บคลอว์เลอร์ตามที่ได้กล่าวมาข้างต้นนี้ ส่วนที่แตกต่างหรือนิยมพัฒนาได้แก่ วิธีในการตรวจสอบว่าเว็บเพจควรมีการเปลี่ยนแปลงไปมากเท่าใดจึงสมควรได้รับการดาวน์โหลด, วิธีจัดลำดับการทำงานของเว็บคลอว์เลอร์ (Scheduler), และความเร็วในการทำงาน ส่วนหลักการทำงานโดยทั่วไปนั้นมีความคล้ายคลึงกันคือ มีการเริ่มต้นการทำงานจากยูอาร์แอลเริ่มต้นแล้วสำรวจและดาวน์โหลดลิงค์ต่างๆของเว็บเพจถูกดาวน์โหลดมาก่อนหน้า

จากการวิเคราะห์พบว่า ลักษณะการรวบรวมข้อมูลจากอินเทอร์เน็ตโดยอาศัยเว็บครอว์เลอร์นั้น ต้องอาศัยระยะเวลานานในการสำรวจไปตามลิงค์ต่างๆในเว็บเพจ อีกทั้งในงานวิจัยนี้ต้องการเฉพาะเว็บเพจที่มีข้อมูลบุคคลเท่านั้น และต้องมีการแปลงเว็บเพจที่อยู่ในรูปแบบของข้อความ ดังนั้นการนำเว็บครอว์เลอร์มาใช้งานเพียงอย่างเดียวจึงไม่เพียงพอต่อความต้องการ ดังนั้นในงานวิจัยนี้ จึงได้นำเสิร์ชเอนจิ้นมาใช้ในการรวบรวมเว็บเพจข้อมูลบุคคลแทน เนื่องจากเสริชเอนจิ้นแต่ละตัวนั้นจะมีเว็บครอว์เลอร์รวมอยู่ด้วย และเสิร์ชเอนจิ้นสามารถสนองตอบความต้องการดังกล่าวได้ ซึ่งในงานวิจัยนี้ได้เลือกใช้เสิร์ชเอนจิ้นของ mnoGoSearch เนื่องจากสามารถรองรับการทำงานกับภาษาไทยได้และเป็นซอฟต์แวร์เสรี เพื่อรวบรวมข้อมูลบุคคลในระบบอินเทอร์เน็ต

<u>วิธีในการตรวจสอบความคล้ายของสตริง</u>

วิธีในการตรวจสอบความคล้ายของสตริงที่ใช้ในในการตรวจสอบว่าสตริง 2 สตริงมีความเหมือนหรือแตกต่างกันมากเพียงใดนั้น คือ การคำนวณค่าความต่างของสตริง (Edit distance) ซึ่งวิธีการนี้คำนวณหาค่าความคล้ายของสตริงจากจำนวนครั้งของการดำเนินการ(operation) ของการเปลี่ยนจากอักขระหนึ่งไปเป็นอีกอักขระหนึ่ง ตัวอย่าง เช่น “ทักษิน” กับ “ทักษิณ” มีส่วนที่ต่างกันคือ “น” และ “ณ” นั่นคือวิธีการหนึ่งของการดำเนินการที่เกิดขึ้น โดยการดำเนินการที่ได้ระบุไว้ในวิธีการตรวจสอบความคล้ายของสตริงดังกล่าวนี้มี 3 แบบ ได้แก่ การแทรก (insertion) การลบ (deletion) และการแทนที่ (substitution) ซึ่งเป็นชื่อที่ตั้งโดยนาย Vladimir Levenshten การนับการดำเนินการที่เกิดขึ้นนี้จะนับจากตัวอักษรทีละตัว วิธีนี้มีประโยชน์มากสำหรับโปรแกรมประยุกต์ที่

ต้องการ การตรวจการสะกดคำ ยกตัวอย่างเช่น ระหว่างคำว่า “ประกิต” กับ “ปะกิทร” มีค่าความแตกต่างที่เกิดขึ้นจากการคำนวณโดยวิธีการคำนวณค่าความต่างเท่ากับ 3 เนื่องจากมีการเปลี่ยนแปลงทั้งหมด 3 ครั้ง ดังนี้

1. “ระกิต” กับ “ะกิต” (มีการลบของ “ร” เกิดขึ้น)
2. “ประกิต” กับ “ประกิท” (มีการเปลี่ยนของ “ต” กับ “ท” เกิดขึ้น)
3. “ประกิต” กับ“ประกิตร” (มีการแทรกของ “ร” เกิดขึ้น)

สำหรับวิธีอื่นที่ใช้ในการวัดความคล้ายของสตริงคือ ฮัมมิ่งดิสแทน (Hamming distance) ซึ่งแตกต่างจากวิธีการแบบการคำนวณค่าความต่างตรงที่ วิธีการนี้จะใช้ในการเปรียบเทียบระหว่างสตริงที่มีความยาวเท่ากันเท่านั้น และพิจารณาการเปลี่ยนแปลงชนิด “การแทนที่” (substitution) เท่านั้น

อัลกอริทึมที่ใช้ในวิธีการคำนวณค่าความต่างนั้นจะใช้การคำนวณจากล่างขึ้นบนของไดนามิคโปรแกรมมิ่ง ในการคำนวณ รวมถึงการใช้ $(n + 1) \times (m + 1)$ เมทริก เมื่อ n และ m คือความยาวของสตริงแต่ละตัว จากภาพที่ 2 เป็นการจำลองการทำงานของวิธีการการคำนวณค่าความต่าง โดยการเปรียบเทียบความคล้ายระหว่างตัวแปร “str1” กับ “str2”

```
int LevenshteinDistance(char str1[1..lenStr1], char str2[1..lenStr2])
   // d is a table with lenStr1+1 rows and lenStr2+1 columns
   declare int d[0..lenStr1, 0..lenStr2]
   // i and j are used to iterate over str1 and str2
   declare int i, j, cost

   for i from 0 to lenStr1
       d[i, 0] := i
   for j from 0 to lenStr2
       d[0, j] := j

   for i from 1 to lenStr1
       for j from 1 to lenStr2
           if str1[i] = str2[j] then cost := 0
                                else cost := 1
           d[i, j] := minimum(
                                d[i-1, j  ] + 1,     // deletion
                                d[i  , j-1] + 1,     // insertion
                                d[i-1, j-1] + cost   // substitution
                            )

   return d[lenStr1, lenStr2]
```

ภาพที่ 2 อัลกอริทึมของวิธีการตรวจสอบความคล้ายของสตริง (Hodge, 2001)

สำหรับงานวิจัยนี้ ได้พัฒนาการคำนวณค่าความแตกต่างของการสะกดคำให้เหมาะสมกับภาษาไทย (ดูรายละเอียดเพิ่มเติมที่หัวข้ออุปกรณ์และวิธีการ)

วิธีที่ใช้ในการจัดเก็บข้อมูล

วิธีที่นิยมใช้ในการจัดเก็บข้อมูลที่เป็นที่นิยมคือ เช่น การจัดเก็บลงในฐานข้อมูล แต่การจัดเก็บโดยวิธีดังกล่าวมีข้อเสียคือ ไม่สะดวกในการใช้งานร่วมกันระหว่างคอมพิวเตอร์ เนื่องจากฐานข้อมูลเป็นการจัดเก็บข้อมูลที่มีความตายตัว ไม่ยืดหยุ่น เช่น ตารางในฐานข้อมูลนั้นมีการระบุชื่อตาราง, มิติของตาราง และชื่อเขตข้อมูลในตารางที่แน่นอน จึงจะสามารถใช้งานฐานข้อมูลได้ ดังนั้นจึงมีความต้องการรูปแบบในการจัดเก็บข้อมูลในลักษณะที่สามารถแบ่งปันและเชื่อมโยงข้อมูลระหว่างคอมพิวเตอร์ร่วมกันได้อย่างสะดวก ซึ่งการจัดเก็บข้อมูลที่มีลักษณะตามความต้องการดังกล่าวนั้นคือ เทคโนโลยีของซีแมนติกเว็บ ซึ่งความรู้พื้นฐานที่เกี่ยวกับซีแมนติกเว็บมีดังต่อไปนี้

ซีแมนติกเว็บ

ซีแมนติกเว็บเป็นส่วนขยายของเว็บในปัจจุบัน ซึ่งทำให้ผู้ใช้สามารถค้นหา , แบ่งปัน เชื่อมโยงความสัมพันธ์ระหว่างข้อมูลได้ง่ายขึ้น เนื่องจากซีแมนติกเว็บถูกสร้างขึ้นโดยใช้ภาษาอาร์ดีเอฟ ซึ่งเป็นภาษาที่ใช้ในการบรรยายข้อมูลให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถทำความเข้าใจได้ โดยผู้ริเริ่มก่อตั้งซีแมนติกเว็บคือนาย Tim Berners-Lee ซึ่งเป็นผู้ริเริ่มก่อตั้ง เวิลด์ ไวด์ เว็บ ในปัจจุบันเช่นกัน

ส่วนประกอบของซีแมนติกเว็บ

ซีแมนติกเว็บนั้นมีพื้นฐานของมาตรฐานอยู่บนภาษาเอ็กซ์เอ็มแอล(XML) โดยภาษาในการจัดเก็บข้อมูลแบบใหม่เพื่อจัดเก็บข้อมูลให้อยู่ในรูปแบบที่เครื่องคอมพิวเตอร์สามารถเข้าใจความหมายได้ (Semantic) ของข้อมูลเหล่านั้นได้ ภาษาเหล่านั้นได้แก่ RDF (Resource Description Framework) และ OWL (Web Ontology Language)

ภาษาอาร์ดีเอฟเป็นโครงสร้างในการจัดเก็บข้อมูลที่ช่วยในการเชื่อมโยงทรัพยากร (Resource) ต่างๆที่มีอยู่ในระบบอินเทอร์เน็ต อย่างมีความสัมพันธ์ ตัวอย่างของทรัพยากรในระบบ

อินเทอร์เน็ตได้แก่ ยูอาร์แอล URL (Uniform Resource Identifier) เช่น http://www.ku.ac.th หรืออีเมลล์แอดเดรส เป็นต้น โดยภาษาอาร์ดีเอฟนั้นมีพื้นฐานอยู่บนภาษา XML และในภาษาอาร์ดีเอฟนั้นจะมีเค้าร่างของอาร์ดีเอฟ (RDF Schema) ซึ่งเป็นที่บรรจุคำศัพท์เพื่อใช้ในการบรรยายคุณสมบัติและคลาสของทรัพยากรอย่างมีความหมาย

ภาษาอาวล์เป็นรูปแบบในการจัดเก็บข้อมูลที่มีลักษณะคล้ายกับภาษาอาร์ดีเอฟเป็นอย่างมาก ข้อแตกต่างระหว่างภาษาอาร์ดีเอฟกับภาษาอาวล์คือ จำนวนของคำศัพท์ โดยภาษาอาวล์จะมีคำศัพท์ที่มากกว่าอาวล์เพื่อใช้ในการบรรยายคุณสมบัติและคลาส ตัวอย่างเช่น

ภาษาอาร์ดีเอฟ (RDF)

RDF ย่อมากจาก Resource Description Framework ซึ่งเป็นภาษาที่มุ่งเน้นเพื่อจะใช้ในการบรรยายหรืออธิบายแหล่งทรัพยากรที่มีอยู่ในระบบอินเทอร์เน็ต โดยรูปแบบที่ใช้บรรยายนั้นจะอยู่ในรูปแบบ ประธาน-ตรรกะความสัมพันธ์-กรรม (subject-predicate-object) ซึ่งถูกเรียกว่าทริปเปิ้ล เนื่องจากมีองค์ประกอบ 3 ส่วน ซึ่ง subject คือรีสอร์ทหรือสิ่งที่จะถูกบรรยาย predicate คือความสัมพันธ์ที่ต้องการใช้บรรยาย สุดท้ายคือ object ซึ่งเป็นค่าที่ความสัมพันธ์นั้นกล่าวอ้างถึงโดยจะอยู่ในรูปแบบของทรัพยากรหรือสตริงก็ได้ ตัวอย่างของการจัดเก็บข้อมูลลงในภาษาแบบอาร์ดีเอฟนั้น แสดงดังภาพที่ 3 และภาพที่ 4

ภาพที่ 3 ตัวอย่างรูปแบบของการบรรยายข้อมูลด้วยภาษาอาร์ดีเอฟ

จากภาพที่ 3 เป็นการบรรยายตัวย่อของรัฐนิวยอร์ก (New York) ในภาษาอังกฤษนั้น ประโยค “New York has a postal abbreviation which is NY” โดยนิวยอร์กจะทำหน้าที่เป็นประธาน Postal abbreviation ทำหน้าที่เป็น predicate และ NY ทำหน้าที่เป็น object เมื่อทำการแปลงประโยคดังกล่าวให้อยู่ในรูปแบบของอาร์ดีเอฟ แล้วจึงได้ประโยคดังแสดงในภาพที่ 4

สำหรับอีกตัวอย่างหนึ่งคือ แหล่งข้อมูลจากวิกิพีเดีย (www.wikipedia.org)ของ Tony Benn ซึ่งในการบรรยายว่าแหล่งข้อมูลนั้นเป็นของนาย Tony Benn และผู้เผยแพร่ข้อมูลคือวิกิพีเดีย แล้วนั้น สามารถเขียนให้อยู่ในรูปแบบของอาร์ดีเอฟได้ดังประโยคต่อไปนี้

<http://en.wikipedia.org/wiki/Tony_Benn> <http://purl.org/dc/elements/1.1/title> "Tony Benn".

<http://en.wikipedia.org/wiki/Tony_Benn> <http://purl.org/dc/elements/1.1/publisher> "Wikipedia"

จากประโยคข้างต้นสามารถเขียนให้อยู่ในรูปแบบอาร์ดีเอฟที่ใช้วากยสัมพันธ์ของเอ็กซ์เอ็มแอลได้ดังภาพที่ 4

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="http://en.wikipedia.org/wiki/Tony_Benn">
  <dc:title>Tony Bena</dc:title>
  <dc:publisher>Wikipedia</dc:publisher>
</rdf:Description>
</rdf:RDF>
```

ภาพที่ 4 การจัดเก็บข้อมูลด้วยโครงสร้างอาร์ดีเอฟโดยใช้วากยสัมพันธ์แบบเอกซ์เอ็มแอล

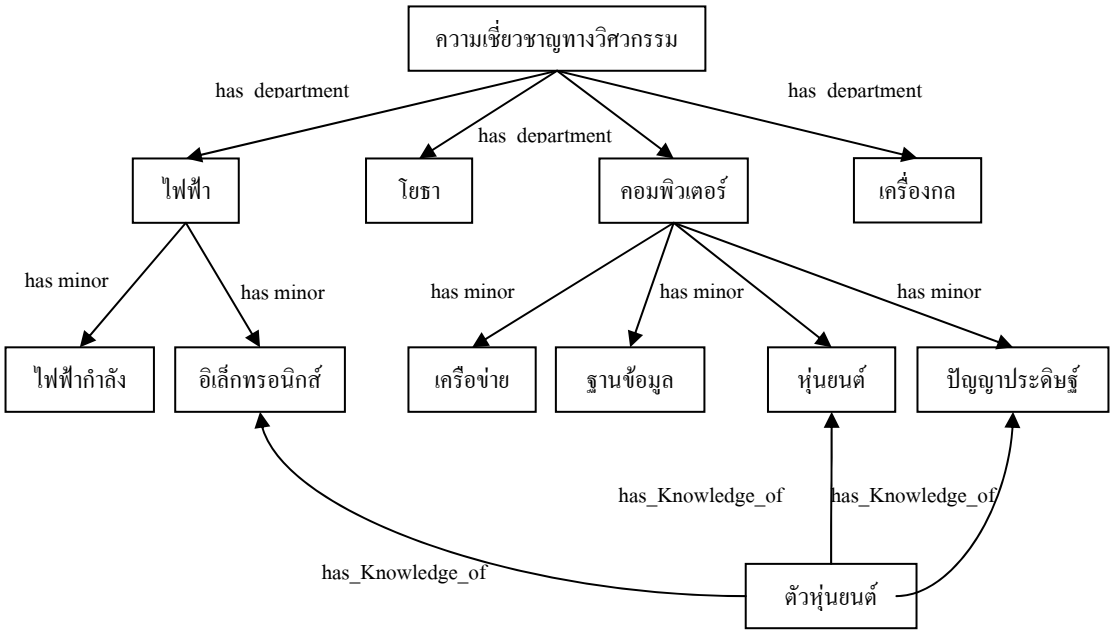
โปรแกรมประยุกต์ที่มีการนำเอาภาษาอาร์ดีเอฟมาประยุกต์ใช้งานได้แก่

- RDF Site Summary ใช้สำหรับการนำข้อมูลสู่สาธารณะเกี่ยวกับ หัวข้อข่าว ข้อมูลที่มีอยู่ในเว็บล็อก (Weblog)
- FOAF (Friend of a Friend) ถูกออกแบบมาเพื่อใช้ในการบรรยายข้อมูลบุคคล ความสนใจของแต่ละบุคคล และความสัมพันธ์ระหว่างบุคคล
- DOAC (Description of Career) เป็นส่วนขยายของ FOAF โดยข้อมูลที่ทำการจัดเก็บคือประวัติการทำงาน
- MusicBrainz นำเสนอข้อมูลเกี่ยวกับอัลบั้มเพลงที่ต้องการนำเสนอสู่สาธารณะ

- Creative Commons ใช้อาร์ดีเอฟเพื่อฝังข้อมูลเกี่ยวกับลิขสิทธิ์หรือการได้รับอนุญาติลงไปในไฟล์เว็บเพจและเอ็มพีทรี (mp3)

ภาษาอาวล์ (OWL)

ภาษาอาวล์ (OWL) ย่อมาจาก Web Ontology Language ซึ่งเป็นภาษาที่ได้รับการพัฒนาต่อจากภาษาอาร์ดีเอฟ (RDF) โดยมีคำศัพท์ที่ใช้ในการบรรยายข้อมูลเพิ่มขึ้นจากเดิมตัวอย่างเช่น OWL: onProperty, OWL: Restriction เป็นต้น ซึ่งภาษาอาวล์นั้นถูกออกแบบมาเพื่อใช้งานกับซีแมนติกเว็บเช่นกัน โดยภาษาอาวล์นั้นประกอบด้วยภาษาย่อยจำนวน 3 ภาษา ได้แก่ OWL Lite, OWL Dl และ OWL full ตัวอย่างของออนโทโลจีของความเชี่ยวชาญด้านวิศวกรรมตามภาพที่ 5 ซึ่งมีความสัมพันธ์ที่ได้กำหนดจำนวนขึ้น 3 ความสัมพันธ์ได้แก่ has_minor, has_department, has_knowledge_of



ภาพที่ 5 ออนโทโลจีของความเชี่ยวชาญด้านวิศวกรรม

เมื่อแปลงออนโทโลจีจากภาพที่ 5ให้อยู่ในรูปของภาษาอาวล์แล้วจะได้ผลลัพธ์ดังภาพที่ 6

```
<?xml version="1.0"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:owl="http://www.w3.org/2002/07/owl#"
   xmlns="http://www.owl-ontologies.com/unnamed.owl#"
  xml:base="http://www.owl-ontologies.com/unnamed.owl">
  <owl:Ontology rdf:about=""/>
  <owl:Class rdf:ID="เครือข่าย">
   <rdfs:subClassOf>
    <owl:Class rdf:ID="คอมพิวเตอร์"/>
   </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="หุ่นยนต์">
   <rdfs:subClassOf>
    <owl:Class rdf:about="#คอมพิวเตอร์"/>
   </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="ตัวหุ่นยนต์">
   <rdfs:seeAlso>
    <หุ่นยนต์ rdf:ID="หุ่นยนต์_19"/>
   </rdfs:seeAlso>
   <rdfs:seeAlso>
    <อิเล็กทรอนิคส์ rdf:ID="อิเล็กทรอนิคส์_18"/>
   </rdfs:seeAlso>
   <rdfs:seeAlso>
    <ปัญญาประดิษฐ์ rdf:ID="ปัญญาประดิษฐ์_17"/>
   </rdfs:seeAlso>
  </owl:Class>
  <owl:Class rdf:ID="ไฟฟ้า">
   <rdfs:subClassOf>
    <owl:Class rdf:ID="ความเชี่ยวชาญทางวิศวกรรม"/>
   </rdfs:subClassOf>
```

```
</owl:Class>
<owl:Class rdf:ID="เครื่องกล">
 <rdfs:subClassOf>
  <owl:Class rdf:about="#ความเชี่ยวชาญทางวิศวกรรม"/>
 </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:about="#คอมพิวเตอร์">
 <rdfs:subClassOf>
  <owl:Class rdf:about="#ความเชี่ยวชาญทางวิศวกรรม"/>
 </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID="โยธา">
 <rdfs:subClassOf>
  <owl:Class rdf:about="#ความเชี่ยวชาญทางวิศวกรรม"/>
 </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID="ไฟฟ้ากำลัง">
 <rdfs:subClassOf rdf:resource="#ไฟฟ้า"/>
</owl:Class>
<owl:Class rdf:ID="ฐานข้อมูล">
 <rdfs:subClassOf rdf:resource="#คอมพิวเตอร์"/>
</owl:Class>
<owl:Class rdf:about="#ความเชี่ยวชาญทางวิศวกรรม">
 <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
 >สาขาความเชี่ยวขาญ</rdfs:comment>
</owl:Class>
<owl:Class rdf:ID="อิเล็กทรอนิคส์">
 <rdfs:subClassOf rdf:resource="#ไฟฟ้า"/>
</owl:Class>
<owl:Class rdf:ID="ปัญญาประดิษฐ์">
 <rdfs:subClassOf rdf:resource="#คอมพิวเตอร์"/>
</owl:Class>
<owl:ObjectProperty rdf:ID="has_department">
 <rdfs:domain rdf:resource="#ความเชี่ยวชาญทางวิศวกรรม"/>
```

```
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >Department in engineer faulty</rdfs:comment>
 </owl:ObjectProperty>
 <owl:DatatypeProperty rdf:ID="has_knowledge_of">
  <rdfs:domain rdf:resource="#ตัวหุ่นยนต์"/>
 </owl:DatatypeProperty>
 <owl:DatatypeProperty rdf:ID="has_minor">
  <rdfs:domain rdf:resource="#ความเชี่ยวชาญทางวิศวกรรม"/>
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >Minor in each major</rdfs:comment>
 </owl:DatatypeProperty>
</rdf:RDF>
```

ภาพที่ 6 ตัวอย่างออนโทโลจีของสาขาความเชี่ยวชาญในภาษาอาวล์

การอนุมาน (Reasoning)

การอนุมานเป็นการให้นำเหตุผลของมนุษย์เพื่อมาใช้สรุปหาข้อมูลเพิ่มเติมจากข้อมูลเดิมที่มีอยู่ ซึ่งการอนุมานนั้นโดยพื้นฐานแล้วมีอยู่ 2 ประเภทคือ ได้แก่

การอนุมานแบบนิรนัย (Deductive reasoning)

คำตอบที่เกิดจากการอนุมานลักษณะนี้เกิดจากการพิจารณาจากตรรกะของข้อมูลที่มีอยู่ และคำตอบของการอนุมานลักษณะนี้จะถูกต้องเสมอ ภาพที่ 7

```
ข้อเท็จจริง : บุคคลที่ทำงานในหน่วยงานของรัฐเป็นข้าราชการ
ข้อเท็จจริง : นายสมชายทำงานในหน่วยงานของรัฐ
บทสรุป :     เพราะฉะนั้น_นายสมชายเป็นข้าราชการ
```

ภาพที่ 7 ตัวอย่างของการอนุมานแบบนิรนัย

การอนุมานแบบอุปนัย (Inductive reasoning)

การอนุมานลักษณะนี้มีความตรงกันข้ามกับการอนุมานแบบนิรนัย เนื่องจากการอนุมานในลักษณะนี้จะทำให้เกิดความรู้ใหม่ขึ้นมา ตัวอย่างเช่น

ข้อเท็จจริง : พระอาทิตย์ขึ้นทางด้านทิศตะวันออกทุกวันจนกระทั่งวันนี้ด้วย
บทสรุป : เพราะฉะนั้นพระอาทิตย์จะขึ้นทางตะวันออกในวันพรุ่งนี้เช่นกัน

การอนุมานแบบ Abductive reasoning

เป็นการอนุมานเพื่อคาดเดาคำตอบที่น่าจะเกิดขึ้น แต่ไม่ยืนยันว่าผลลัพธ์ที่เกิดขึ้นจากการอนุมานนั้นจะถูกต้องเสมอไปหรือไม่ ตัวอย่างเช่น

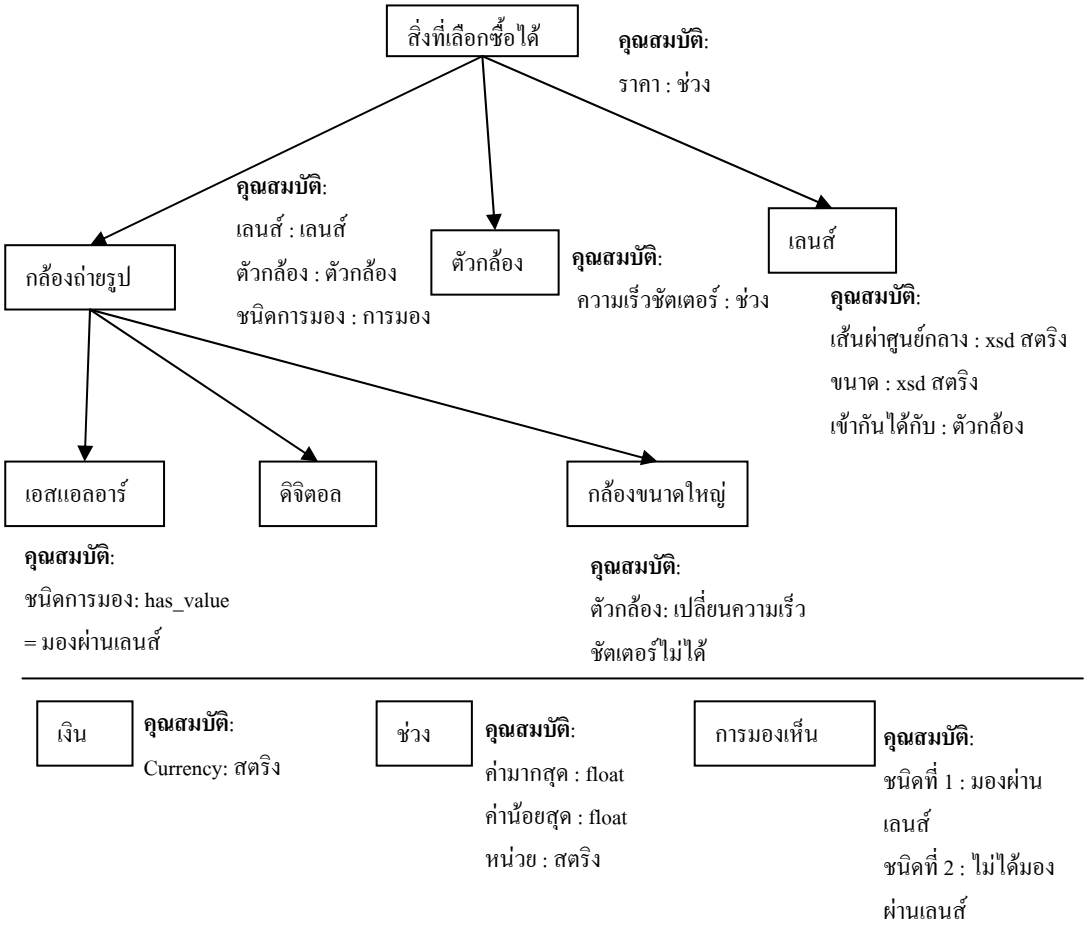
ข้อเท็จจริง: ถ้าเป็นไข้แล้วจะตัวร้อน ข้อเท็จจริง: มีอาการตัวร้อน
บทสรุป: น่าจะเป็นไข้

ออนโทโลจี (Ontology)

ออนโทโลจีเป็นโครงสร้างของการจัดเก็บข้อมูล โดยการจัดเก็บนั้นมุ่งเน้นเพื่อจัดเก็บข้อมูลเป็นกลุ่มๆ ซึ่งคำภายในกลุ่มนั้นจะถูกเชื่อมโยงความสัมพันธ์ที่เป็นไปได้เข้าด้วยกัน การใช้งานออนโทโลจีนั้นโดยส่วนมากจะใช้ใน สาขาวิชาของวิทยาศาสตร์คอมพิวเตอร์ , ปัญญาประดิษฐ์, ซีแมนติกเว็บ เป็นต้น ซึ่งโดยทั่วไปแล้วออนโทโลจีจะประกอบด้วยองค์ประกอบดังต่อไปนี้

- คอนเซ็ปต์ (concept): วัตถุและกลุ่มของวัตถุ (เป็นที่รู้จักกันดีในนามของ คลาส หรือ กลุ่ม
- ลักษณะของวัตถุ (characteristics) : เช่น slots , roles หรือ ontology
- ความสัมพันธ์ (relations) : รูปแบบของการแสดงว่าวัตถุหนึ่งๆ จะมีความสัมพันธ์กับวัตถุอื่นได้อย่างไร

ตัวอย่างของออนโทโลจี



ภาพที่ 8 ออนโทโลจีของกล้องถ่ายรูป (Costello, 2001)

จากภาพที่ 8 เป็นออนโทโลจีที่ใช้จัดเก็บความรู้เกี่ยวกับกล้อง ซึ่งจะระบุกล้องมีส่วนประกอบอะไรบ้าง และส่วนประกอบแต่ละชิ้นนั้นมีความหมายว่าอย่างไร เช่น กล้องมี 3 ประเภทคือ เอสแอลอาร์, ดิจิตอล, และ กล้องขนาดใหญ่ โดยกล้องประเภท เอสแอลอาร์ นั้นมีคุณสมบัติคือ เป็นกล้องที่ใช้แบบมองผ่านเลนส์ โดยสังเกตุได้จาก การมองเห็นชนิดการมอง: has_value = มองผ่านเลนส์ ตามภาพที่ 8

ชนิดของออนโทโลจี

ออนโทโลจีเฉพาะด้าน (Domain ontology)

ออนโทโลจีนี้จะนำเสนอข้อมูลเฉพาะทาง ยกตัวอย่างเช่น คำว่า “การ์ด” สามารถมีได้หลากหลายความหมาย ภายในออนโทโลจีของการเล่นโป๊กเกอร์ คำว่า “การ์ด” จะหมายถึง “ไพ่” แต่ถ้าไปอยู่ในออนโทโลจีของฮาร์ดแวร์คอมพิวเตอร์แล้วจะหมายถึง การ์ดวีดีโอ หรือ การ์ดเสียง

อีกตัวอย่างหนึ่งคือ ข้อมูลกำกับเอกสารหรือหนังสือในห้องสมุด ใช้ Dublin Core เป็นออนโทโลจี ตัวอย่างเช่น ผู้แต่ง, ชื่อหนังสือ, เป็นต้น

ออนโทจีแบบทั่วไป (General ontology)

ใช้ในการบรรยายสรรพสิ่งต่างที่มีอยู่ในโลกโดยทั่วไป ยกตัวอย่างเช่น สัตว์ ประกอบด้วยสัตว์บก สัตว์ปีก สัตว์น้ำ สัตว์ครึ่งบกครึ่งน้ำ เป็นต้น

ภาษาสำหรับสร้างออนโทโลจี

- Cycl เป็นภาษาที่ถูกพัฒนาขึ้นโดยมีพื้นฐานอยู่บน first-order predicate calculus (http://cyc.com/)
- KIF (Knowledge Interchange Format)ใช้วากยสัมพันธ์แบบ first-order logic (http://logic.stanford.edu/kif/kif.html)
- อาวล์ (OWL) เป็นภาษาสำหรับใช้บรรยายความหมายของออนโทโลจีโดยเฉพาะ ซึ่งมีการพัฒนาต่อจากภาษาRDF และ RDFS ซึ่งเป็นภาษาที่ดีใกล้เคียงกับสองภาษาก่อนนี้ ดังตัวอย่างที่แสดงในภาพที่ 6

ตัวอย่างจริงของออนโทโลจี

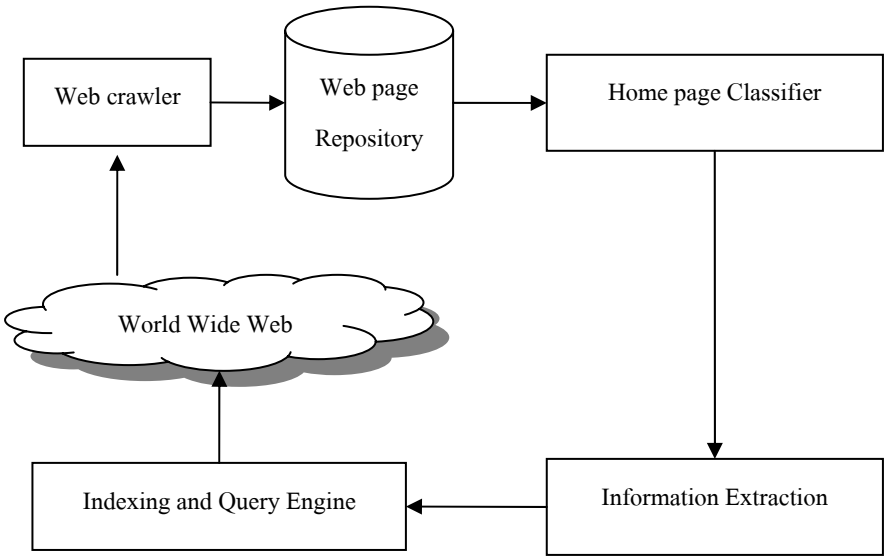
- Wordnet เป็นพจนานุกรมภาษาอังกฤษที่ถูกจัดเก็บในแบบออนโทโลจี
- Gene ออนโทโลจี ใช้ในการเก็บความรู้เกี่ยวกับความสัมพันธ์ระหว่างยีน
- Plant ออนโทโลจี ใช้เก็บความรู้ของโครงสร้างพืช และวิธีการปลูก

งานวิจัยที่เกี่ยวข้อง

ตารางที่ 1 งานวิจัยเกี่ยวกับระบบสืบค้นข้อมูลบุคคลที่ผ่านมา

ปี	งานวิจัย	ข้อมูลที่ใช้ทดลอง	ปัญหาที่สนใจ	ส่วนประกอบระบบ
2002	Searching People on the Web According to Their Interests (Liu, 2002)	โฮมเพจส่วนบุคคลจากหน่วยงานมหาวิทยาลัยจำนวน 10มหาวิทยาลัย	การสกัดข้อมูลของบุคคล, หัวข้องานวิจัยที่สนใจ และหัวข้อการสอนที่สนใจ ออกจากโฮมเพจของแต่ละบุคคลเพื่อเปิดให้บริการในการสืบค้น	-Web Crawler, -Homepage Classifier, -Information Extraction, -Indexing and Query Engine
2003	Mining Web Sites Using Unsupervised Adaptive Information Extraction (Dingli, 2003)	โฮมเพจส่วนบุคคลของบุคคลที่ทำงานที่องค์กรต่างๆซึ่งมีข้อมูลอยู่บนเว็บเพจ	การลดการกำกับข้อมูลบนเว็บเพจด้วยบุคคลโดยใช้ตัวอย่างที่ได้กำกับไว้แล้วบางส่วนเพื่อสอนระบบให้กำกับข้อมูลได้เอง	-Identifying lists of relevant names -Retrieving personal and projects'pages -Extraction of personal data, -Identifying communities of pratice
2004	Extracting social networks and contact information from email and the web (Culotta, 2004)	อีเมลล์จำนวน 49 อีเมลล์และโฮมเพจส่วนบุคคลจำนวน 229 โฮมเพจ	การสกัดความสัมพันธ์ระหว่างบุคคลในสังคมและข้อมูลการติดต่อจากอีเมลล์และเว็บไซต์ โดยวิธีในการสกัดข้อมูลที่ใช้นั้นจะใช้วิธีการสกัดบนพื้นฐานของสถิติและการเรียนรู้	-Person name extraction -Name Co reference -Homepage retrieval -Contact Information and Person name extraction -Keyword extraction -Social Network Analysis
2004	Finding Authoritative People from the Web (Harada, 2004)	เว็บเพจที่มีข้อมูลบุคคลปรากฏอยู่	การสกัดชื่อเฉพาะที่มีอยู่ในเว็บเพจ	Extracting name
2005	Person Resolution in Person Search Results: WebHawk (Wan, 2005)	ข้อมูลบุคคล 200 Rank แรกจาก 200 log ใน msn	การสกัดข้อมูลบุคคล	Filter , Extractor, Cluster
2005	A Testbed for People Searching Strategies in the WWW (Artiles, 2005)	เว็บเพจ	แก้ปัญหาความกำกวมของชื่อคนและการรวบข้อมูลจากหลายแหล่งข้อมูลของบุคคลคนเดียวกันเข้าไว้ด้วยกัน	เว็บครอว์เลอร์

Bing Liu และ Chee Wee Chin ในปี 2003 นำเสนองานวิจัยที่มีชื่อว่า Searching People on the Web According to Their Interest (Bing Liu, 2003) ซึ่งเป็นการพัฒนาระบบสืบค้นข้อมูลที่มีชื่อว่า BullsI Search โดยให้บริการในการสืบค้นข้อมูลในบางโดเมนเท่านั้น ตัวอย่างเช่น ทำให้ผู้ใช้สามารถหาโฮมเพจ, อีเมลล์แอดเดรส ของสมาชิกในคณะวิทยาศาสตร์คอมพิวเตอร์ได้ โดยจุดประสงค์หลักของการทำงานของ BullsI Search นี้คือ การค้นหานักวิจัย โดยใช้ ชื่องานวิจัย หรือชื่อหัวข้อการสอนที่สนใจเป็นคำค้น ซึ่งในการทดลองจะใช้ข้อมูลเช่น ชื่อหัวข้อวิจัยที่สนใจ, อีเมลล์แอดเดรส ที่สกัดได้มาจากกลุ่มของโฮมเพจของสมาชิกของคณะวิทยาศาสตร์คอมพิวเตอร์ ซึ่งมีจำนวน 644 โฮมเพจจากประเทศ สหรัฐอเมริกา,อังกฤษ, แคนาดา, ออสเตรเลีย, นิวซีแลนด์, อิสราเอล, ฮองกง และสิงคโปร์ โดยมีสมาชิกจำนวนคนทั้งสิ้น 25,000 คน แบบจำลองขั้นตอนการทำงานของระบบ BullsI Search มีดังต่อไปนี้



ภาพที่ 9 แบบจำลองระบบสืบค้นข้อมูลบุคคลของ BullsI Search (Bing Liu, 2003)

สำหรับขั้นตอนในการทำงานนั้นเริ่มต้นจาก web crawler ไปรวบรวมเว็บเพจจากคณะวิทยาศาสตร์คอมพิวเตอร์มาไว้รวมกัน โดยได้มาจาก Google web directory (www.google.com) และ yahoo web directory (www.yahoo.com) หลังจากนั้นระบบจะตรวจสอบว่าหน้าใดเป็นหน้าโฮมเพจ โดยตรวจสอบที่ URL ว่ามีเครื่องหมาย “?” หรือไม่ ซึ่งถ้ามีแสดงว่าหน้านั้นไม่ใช่โฮมเพจ หลังจากนั้น HTML tag จะถูกกำจัดออก เพื่อส่งเว็บเพจไม่สู่ส่วนสกัดข้อมูล (Information Extraction) ต่อไป สำหรับวิธีการสกัดข้อมูลจะถูกแบ่งตามประเภทของข้อมูลออกเป็น 2 วิธีคือ

การสกัดความสนใจในงานวิจัยและการสอน (Research and Teaching interest)

ในการสกัดข้อมูลลักษณะนี้ออกจากเว็บเพจประกอบด้วยขั้นตอน 3 ขั้นตอนคือ 1) คัดแยกส่วนที่มีเนื้อหาเกี่ยวกับงานวิจัยและการสอนที่สนใจออกจากหน้าโฮมเพจ 2) ตรวจสอบหาลิงค์ที่นำไปสู่หน้าเว็บเพจที่มีเนื้อหาเกี่ยวกับงานวิจัยและการสอน 3) ใช้การวิเคราะห์ทางภาษาศาสตร์เพื่อสกัดเอาข้อมูลออกมา ตัวอย่างเช่น “My research interest includes…”, “I am teaching …”

การสกัดชื่อคนและอีเมลล์แอดเดรส (Email addresses and names)

สำหรับขั้นตอนในสกัดข้อมูลชื่อคนและอีเมลล์แอดเดรสนั้นมีพื้นฐานความรู้ที่ได้มาจากผู้ทำวิจัย 2 ประการคือ 1) อีเมลล์ของสมาชิกของคณะจะมีความคล้ายกับโฮมเพจ URL มาก 2) อีเมลล์แอดเดรสไอดี (อักขระที่อยู่หน้าเครื่องหมาย @ ในอีเมลล์) จะคล้ายกับชื่อเจ้าของโฮมเพจมาก ดังนั้นวิธีในการสกัดอีเมลล์คือ ในอักขระ @ เพื่อตรวจหาส่วนที่เป็นอีเมลล์ แล้วจึงใช้วิธีการของ Levenshtein Distance เพื่อตรวจสอบความคล้ายระหว่างอีเมลล์ที่มีอยู่กับ URL โฮมเพจ สำหรับอีเมลล์ที่ไม่คล้ายกับ URL โฮมเพจจะไม่ถูกสกัดออกมา ส่วนการสกัดชื่อคนจะดูจากอีเมลล์ที่สกัดมาได้แล้วระบุว่าส่วนที่อยู่หน้าเครื่องหมาย “@” คือชื่อคน

การทำอินเด็กซ์และการสืบค้น(Indexing and Querying Engine)

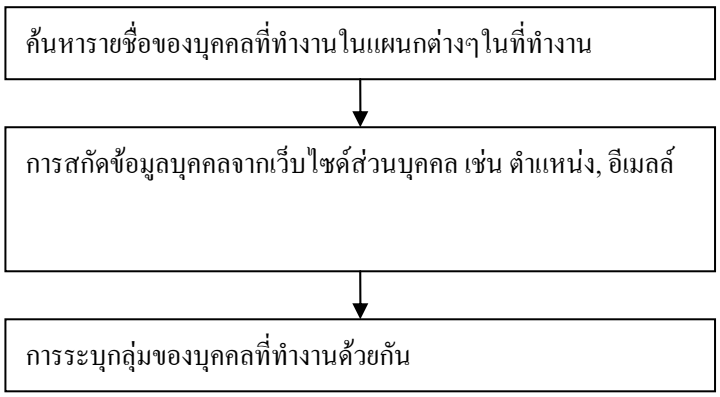
ส่วนนี้เป็นขั้นตอนสุดท้ายของระบบคือการทำอินเด็กซ์(index) จากข้อมูลที่สกัดมาได้เพื่อเตรียมไปสู่การสืบค้น โดยในส่วนนี้จะ MG [Managing Gigabytes: Compression and Indexing Documents and Images] เพื่อทำอินเด็กซ์ สำหรับการสืบค้นข้อมูลจากระบบ BullsI Search มีการออกแบบ Query Interface ดังนี้

ภาพที่ 10 ส่วนติดต่อกับผู้ใช้ของระบบ BullsI Search (Bing Liu, 2003)

ประสิทธิภาพการทำงานของระบบ BullsI Search

Bulls Search ใช้ Precision และ Recall ในการวัดประสิทธิภาพของการสกัดข้อมูล ในการทดลองนั้นเริ่มต้นด้วยการสกัดข้อมูลออกจากเว็บไซต์ของมหาวิทยาลัยจำนวน 10 มหาวิทยาลัย ผลของการสกัด มีดังต่อไปนี้ การสกัดอีเมลล์ (ความความถูกต้อง = 99.03%และ ค่าระลึกได้ = 96.38%), Name Extraction (ค่าความถูกต้อง = 97.51%และ ค่าระลึกได้ =97.38%), การสกัดความสนใจใน (ค่าความถูกต้อง = 95.04% และ ค่าระลึกได้ =95.63%), ความสนใจด้านการสอน (ค่าความถูกต้อง = 90.81%และ ค่าระลึกได้ =85.45%)

Alexiei Dingli, Fabio Ciravegna, David Guthrie และ Yorick Wilks, นำเสนองานวิจัย Mining Web Sites Using Unsupervised Adaptive Information Extraction (Dingli, 2003) ในงานนี้มุ่งเน้นการลดการกำกับเว็บเพจ (Manual annotation) โดยมนุษย์ ซึ่งจะใช้ตัวอย่างของเอกสารที่ได้กำกับไว้แล้วจำนวนหนึ่งเพื่อให้ระบบเรียนรู้ในการกำกับเว็บเพจได้เอง (Bootstrap learning) ระบบสามารถที่สร้างรูปแบบในการกำกับเอกสารเพิ่มขึ้นได้เองหลังจากที่ได้เริ่มทำการกำกับเอกสาร ขั้นตอนในการทำงานของระบบแสดงในภาพที่ 11



ภาพที่ 11 แบบจำลองการทำงานของการสืบค้นข้อมูลบุคคลในองค์กร (Dingli, 2003)

จากภาพที่ 11 ข้อมูลที่ใช้ในการทดลองคือ เว็บไซด์ของคณะวิทยาการคอมพิวเตอร์เพื่อใช้ในการสกัดข้อมูลเกี่ยวกับบุคคลและค้นหากลุ่มของการทำงาน เช่น ใครทำงานอยู่กับใคร เป็นต้น ในขั้นตอนสุดท้ายของการทำงานจากภาพที่ 11 นั้น การระบุกลุ่มของบุคคลที่ทำงานด้วยกันตรวจสอบได้จาก งานวิจัยที่ทำร่วมกัน และสิ่งตีพิมพ์ที่มีชื่อร่วมกัน คำอธิบายการทำงานในแต่ละขั้นตอนการทำงานมีดังต่อไปนี้

การหารายชื่อของบุคคลกลุ่มเดียวกัน (Identifying lists of relevant names)

ใช้วิธี Name Entity Recognizer ที่มีชื่อว่า NERC ในการระบุชื่อบุคคล โดยวิธีการนี้ให้ความถูกต้องมากกว่า 90% ชื่อที่สกัดมาได้นั้นจะถูกส่งไปว่าเป็นชื่อที่มีอยู่จริงหรือไม่ที่ CiteSeer (www.citeseer.com) การส่งชื่อไปตรวจสอบที่ Citeseer นั้นจะทำให้ได้รายชื่อของบุคคลที่มีความเกี่ยวข้องกับชื่อที่ถูกส่งไปตรวจสอบด้วย เช่น เป็นผู้แต่งสิ่งตีพิมพ์ร่วม (Co-author)

การรวบรวมเว็บส่วนบุคคลและโครงงานวิจัย (Retrieving personal and project' pages)

ในการค้นหาโฮมเพจของโครงงานวิจัยและโฮมเพจส่วนบุคคลโดยใช้ Classifiers และ named entity recognizers โดยในการเริ่มต้นการทำงานของวิธีการเหล่านี้ชื่อของบุคคลและโครงงานวิจัยจำนวนหนึ่งเพื่อเป็นค่าเริ่มต้นสำหรับ classifiers และ named entity recognizers อีกทั้งยังใช้วิธีการค้นหาโฮมเพจของโครงงานวิจัยและโฮมเพจส่วนบุคคลจาก hypertext link บริเวณที่อยู่ใกล้ๆกับชื่อบุคคลที่สกัดมาได้ หรือค้นหาจากเสิร์ชเอนจิ้น

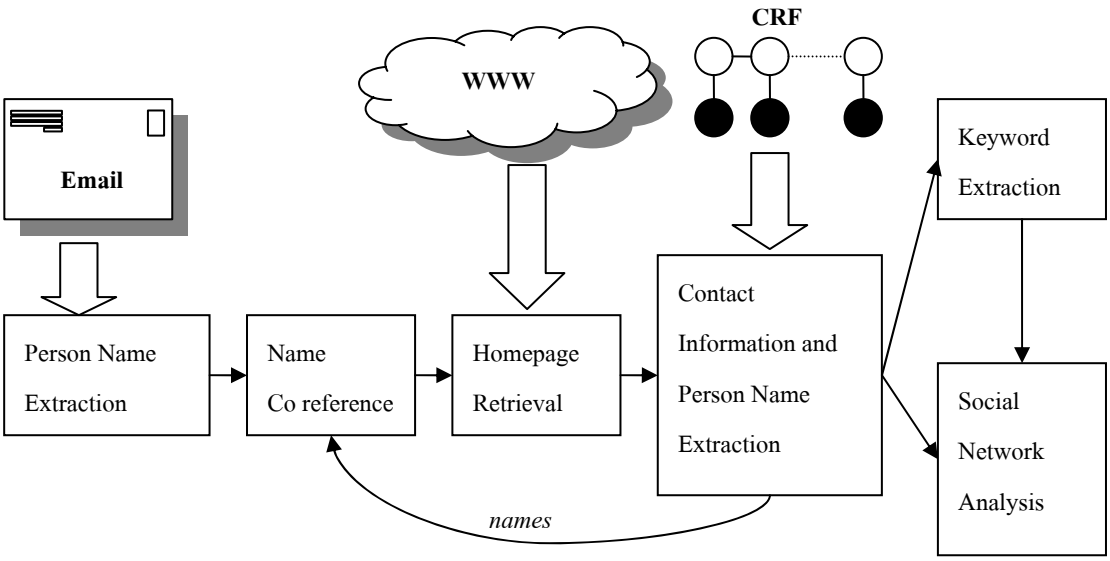
การสกัดข้อมูลบุคคล (Extraction of personal data)

ใช้ wrapper (Kushmarick et al., 1997) ซึ่งเป็นวิธีในสกัดเฉพาะข้อมูลที่ต้องการออกจากเว็บเพจที่มีรูปแบบหนึ่งๆ อย่างอัติโนมัติ โดยผู้ทำการสกัดข้อมูลจะต้องรู้รูปแบบของการปรากฏของข้อมูลที่ต้องการ หลังจากนั้นจึงส่งเว็บเพจที่มีข้อมูลที่ต้องการไปให้ wrapper เพื่อเป็นตัวอย่างสำหรับการเรียนรู้รูปแบบการสกัด ซึ่งเว็บเพจเหล่านั้นจะมีรูปแบบที่มีลักษณะคล้ายกัน เมื่อผู้ใช้ระบุรูปแบบในการสกัดข้อมูลให้แก่ wrapper แล้ว wrapper ก็จะทำการสกัดเว็บเพจอื่นๆที่มีรูปแบบคล้ายๆกันอย่างอัตโนมัติต่อไป

การหากลุ่มข้อมูลที่สอดคล้องกัน (Identifying communities of practice)

ในการจัดกลุ่มคนที่มีความเกี่ยวข้องกัน ใช้ข้อมูลจากห้องสมุดดิจิทัล (Digital library) ที่มีอยู่ ในการสร้างรายการของ คำนำหน้าบุคคล ผู้แต่งหนังสือร่วม และ วันที่ตีพิมพ์ เพื่อใช้เป็นรูปแบบในการเรียนรู้สำหรับ Wrapper ในการสกัดข้อมูลที่เกี่ยวข้องมาเพิ่มในรายการที่สอดคล้องกัน

Aron Culotta, Ron Bekkerman and Andrew McCallum นำเสนองานวิจัย Extracting social networks and contact information from email and the Web (Culotta, 2004) งานวิจัยนี้นำเสนอระบบซึ่งทำหน้าที่สกัดความสัมพันธ์ระหว่างบุคคลในสังคมและข้อมูลการติดต่อจากอีเมลล์และเว็บไซด์ โดยวิธีในการสกัดข้อมูลที่ใช้นั้นจะใช้วิธีการสกัดบนพื้นฐานของสถิติและการเรียนรู้ (Statistics- and learning-based) ซึ่งกระบวนการทำงานหลักๆของระบบมีดังต่อไปนี้



ภาพที่ 12 แบบจำลองการทำงานของระบบสืบค้นข้อมูลบุคคล (Culotta, 2004)

ในการออกแบบนั้นผู้วิจัยได้มุ่งหวังให้ระบบมีความสามารถดังต่อไปนี้คือ 1) ค้นหาผู้เชี่ยวชาญที่สังกัดในบริษัทเอกชนและกลุ่มผู้ทำงานด้านวิทยาศาสตร์ 2) วิเคราะห์กราฟที่แสดงความสัมพันธ์ทางสังคมของข้อมูลคนที่สกัดได้เพื่อค้นหาบุคคลที่มีความเชี่ยวชาญหรือเป็นที่รู้จักในด้านนั้นๆ 3) ค้นหาข้อมูลสำหรับติดต่อไปยังกลุ่มธุรกิจหรือสังคมซึ่งสอดคล้องกับสิ่งที่ผู้ใช้ต้องการ 4) แบ่งแยกกลุ่มคนโดยดูจากกราฟที่แสดงการติดต่อสื่อสาร โดยจากความต้องการของระบบที่กล่าวมานั้นถูกออกแบบเป็นระบบดังภาพที่ 12

จากภาพที่ 12 อินพุตของระบบคือ ข้อความอีเมลล์ที่อยู่ในกล่องรับของความของผู้ใช้ และเอาท์พุต (Output) ของระบบคือ การเติมข้อมูลของบุคคลและข้อมูลการติดต่อกับคนๆนั้นเข้าไปยังแอดเดรสบุค (address book) ของผู้ใช้แต่ละคนแบบสอดคล้องกับความสัมพันธ์ที่ผู้ใช้ได้ระบุไว้ในตอนต้น ซึ่งจากภาพที่ 12 การทำงานหลักมีดังต่อไปนี้

การสกัดชื่อบุคคล (Person name extraction)

สกัดจากส่วนหัวของข้อความในอีเมลล์โดยใช้รูปแบบ(Pattern) ที่ได้ศึกษาไว้ในการสกัดชื่อคน

การจัดการชื่อที่เขียนได้หลายรูปแบบ (Name coreference)

ส่วนนี้เป็นการรวมชื่อคนที่เป็นบุคคลคนเดียวกันแต่เขียนไม่เหมือนกันเข้าด้วยกันเช่น “Joseph Conrad” กับ “J Conrad” สำหรับการแก้ปัญหานี้นั้นจะใช้ String matching rules ที่ได้พัฒนาขึ้นมาเอง

การเสาะหาโฮมเพจ (Homepage retrieval)

หลังจากผ่านสองขั้นตอนแรกมากแล้วก็จะได้เซ็ตของชื่อคนจำนวนหนึ่ง หลังจากนั้นชื่อคนจะถูกส่งไปเป็นคำค้นใน Google search engine เพื่อค้นหาโฮมเพจของคนๆนั้นตัวอย่างเช่นถ้าชื่อคือ Tom Mitchell คำค้นที่ถูกส่งไปจะเป็น (“Tom Mitchell site:cs.cmu.edu”) ซึ่ง “cs.cmu.edu” เป็นค่าที่ได้มาจากอีเมลล์แอดเดรส เมื่อผ่านขั้นตอนนี้ก็จะได้เซ็ตของ URL มาจำนวนหนึ่ง สิ่งที่ต้องทำต่อไปคือการกรองให้เหลือเฉพาะ URL ที่เป็นโฮมเพจเท่านั้น โดยใช้วิธี String kernel distance measure [Text classification using string kernels] ซึ่งเป็นการวัดความคล้ายกันของสตริง(string) โดยเปรียบเทียบระหว่างชื่อคนกับบางส่วนของ URL ซึ่งถ้าเป็นโฮมเพจแล้วสองส่วนนี้จะมีความคล้ายกัน สุดท้ายแล้วก็จะได้โฮมเพจของแต่ละคน

การสกัดชื่อบุคคลและข้อมูลการติดต่อ (Contact information and person name extraction)

ใช้ Probabilistic information extraction model เพื่อค้นหาชื่อของบุคคลและข้อมูลการติดต่อจากโฮมเพจที่ได้มาจากขั้นตอนก่อนหน้านี้ หลังจากนั้นระบบจะสร้าง social network ขึ้นมาซึ่งเป็นการเชื่อมโยงระหว่างบุคคลและเจ้าของเว็บเพจที่บุคคลนี้ถูกค้นพบ

การสกัดคำสำคัญ(Expertise keyword extraction)

การสร้าง Keyword เพื่อเป็นตัวแทนของแต่บุคคลโดยวิเคราะห์จากเทอม (term) ที่ได้มาจากโฮมเพจส่วนบุคคล (Personal Homepage) ที่ทำให้คนๆนั้นต่างจากบุคคลอื่นมากที่สุด

การวิเคราะห์เครือข่ายทางสังคม (Social network)

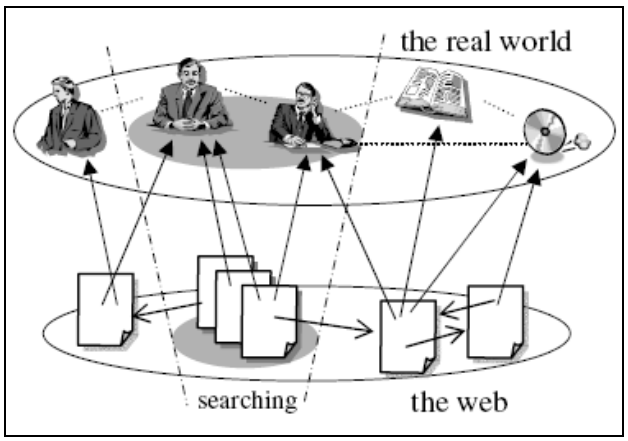
ผลลัพธ์ของเครือข่ายทางสังคมคือกลุ่มของบุคคลที่มีความสัมพันธ์กัน ซึ่งวิธีในการแบ่งเพื่อให้ได้กลุ่มของบุคคลเหล่านี้นั้น จะใช้อัลกอริทึมของ graph partitioning ซึ่งจะทำหน้าที่วิเคราะห์หาโหนด (แทนด้วยบุคคล) ที่มีการเชื่อมต่อเป็นจำนวนมาก แล้วตัดแยกให้เป็นกลุ่มอิสระเพื่อแยกกลุ่มเครือข่ายทางสังคม

ประสิทธิภาพการทำงานของงานวิจัยนี้

ทดลองกับบุคคลที่เป็นสมาชิกในโครงการ CALO (Cognitive Assistant that Learns and Organizes.)จำนวน 2 คน ซึ่งข้อความในอีเมลล์ส่วนใหญ่มีความสอดคล้องกับ CALO project และผู้รับข้อความอีเมลล์ก็เป็นสมาชิกใน CALO project ด้วย ผลการทดลองมีดังนี้ สำหรับผู้ใช้คนแรกมีข้อความในอีเมลล์จำนวน 644 ข้อความ หลังจากที่สกัดชื่อของบุคคลออกจากส่วนหัวของอีเมลล์และได้ social network ที่เชื่อมโยงระหว่างบุคคลจำนวน 53 การเชื่อมต่อ ส่วนข้อมูลของผู้ถูกทดลองคนที่ 2 มีอีเมลล์จำนวน253 ฉบับ และมี social network จำนวน 49 การเชื่อมต่อเครือข่าย

Masanori Harada , Shin-ya Sato และ Kazuhiro Kazama นำเสนองานวิจัย Finding Authoritative People from the Web (Harada, 2004)งานวิจัยนี้นำเสนอ NEXAS (Name Entity eXtraction and Association Search) ซึ่งเป็นส่วนขยายของระบบเสิร์ชเอนจิ้น (Search Engine) โดยทำหน้าที่ในการสกัดชื่อเฉพาะ (Proper name) ที่มีอยู่ในเว็บเพจ และชื่อเฉพาะที่ NEXAS สกัดออกมานั้นจะมีความสอดคล้องกับคำค้นที่ผู้เลือกใช้ซึ่งประโยชน์ที่สำคัญจากการใช้งานระบบนี้คือผู้ใช้สามารถสืบค้นข้อมูลในหัวข้อใดๆก็ได้ที่สนใจหลังจากนั้นผู้ใช้จะได้รับข้อมูลเกี่ยวกับชื่อเฉพาะที่เกี่ยวข้องกับหัวข้อนั้นๆ โดยในการทดลองครั้งนี้ได้มุ่งเน้นไปยังการค้นหาบุคคลที่เป็นผู้แต่งหนังสือจากเว็บเพจ ซึ่งเป็นประโยชน์มากในทางปฏิบัติ เนื่องจากเมื่อผู้ใช้ได้ชื่อของผู้แต่งหนังสือที่เกี่ยวข้องกันหัวข้อที่ค้นหาจากเว็บเพจแล้ว ผู้ใช้สามารถนำชื่อที่ได้มานั้นไปสืบค้นเพิ่มเติมในห้องสมุดหรือแหล่งสืบค้นใดๆได้อีกด้วย โดยที่ไม่จำเป็นต้องเว็บที่ได้มาจากการสืบค้นทีละหน้าเพื่อสกัดเอาชื่อของบุคคลที่เกี่ยวข้องด้วยตัวเอง

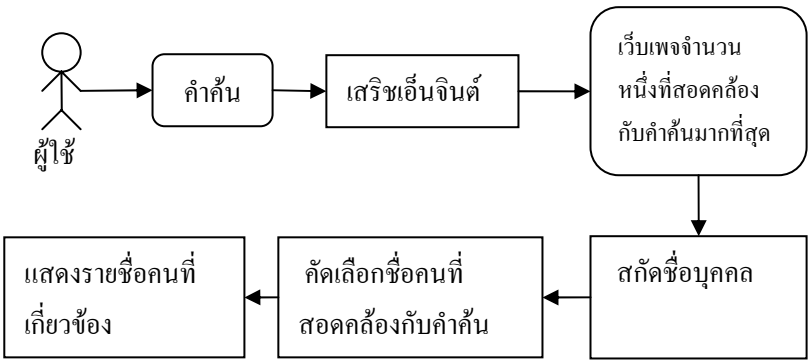
เป้าหมายโดยรวมของระบบ NEXAS นี้ คือการพยายามจำลองความสำพันธ์ระหว่างข้อมูลที่มีอยู่ในระบบอินเทอร์เน็ต ไปสู่วัตถุที่มีปรากฏอยู่จริงในโลก ดังแสดงใน ภาพที่ 13



ภาพที่ 13 ความสัมพันธ์ระหว่างข้อมูลของระบบ NEXAS (Harada, 2004)

ซึ่งสุดท้ายแล้วผู้วิจัยมุ่งหวังที่จะให้ระบบสามารถตอบคำถามเหล่านี้ได้ “ใครมีความเกี่ยวข้องกับหัวข้อ (คำค้น) นี้”, “ใครเป็นบุคคลผู้ซึ่งเป็นที่เป็นที่รู้จักมากที่สุดในโลก”, “ในขณะนี้เพลงใดได้รับนิยมมากที่สุด”, “หนังสือเล่มไหนที่มีความเกี่ยวข้องกับหนังสือเล่มนี้ (คำค้น) บ้าง” เป็นต้น แต่ในขณะนี้การทดลองจะทำเพียงแค่ค้นหาผู้แต่งหนังสือจากหัวข้อตามคำค้นที่ผู้ใช้เลือกใช้เท่านั้น

NAXAS เป็นระบบที่สกัดชื่อคนที่สอดคล้องกับคำค้นที่ผู้ใช้เลือกใช้ โดยสกัดมาจากลำดับต้นๆของเว็บเพจที่เป็นผลลัพธ์จากการสืบค้นโดยเสิร์ชเอนจิ้น โดยสมมุติฐานเบื้องต้นในการสกัดข้อมูลคือ ชื่อคนที่สอดคล้องกับคำค้นที่ผู้ใช้เลือกใช้คือ ชื่อคนที่ประกฏซ้ำในเว็บเพจที่เป็นผลลัพธ์ซึ่งได้มาจากการค้นคืนของเสิร์ชเอนจิ้น โดยแบบจำลองการทำงานของ NAXAS แสดงในภาพที่ 14



ภาพที่ 14 แบบจำลองการทำงานระบบสืบค้นข้อมูลบุคคลของ NEXAS (Harada, 2004)

จากภาพที่ 14 ขั้นตอนการทำงานหลักของระบบ NAXAS มีดังต่อไปนี้

<u>การรวบรวมแหล่งข้อมูล</u>

ใช้เสิร์ชเอนจิ้นเพื่อค้นคืนเว็บเพจที่สอดคล้องกันคำค้นโดยจะเลือกเว็บเพจที่อยู่ในลำดับต้นของการค้นคืนออกมาจำนวนหนึ่งเพื่อนำไปใช้ในการสกัดเอาชื่อคนที่เกี่ยวข้องกับคำค้นออกมา ซึ่งในการทดลองครั้งนี้ได้ใช้เสิร์ชเอนจิ้นที่มีชื่อว่า ODIN (A searching and ranking scheme using hyperlinks and anchor text, Social network analysis) ซึ่งเป็นเสิร์ชเอนจิ้นที่ทางทีมวิจัย NAXAS ได้พัฒนาขึ้นมา

<u>การสกัดชื่อบุคคล</u>

ในการสกัดชื่อบุคคลออกจากเอกสารใดๆนั้นสามารถกระทำได้หลายวิธีเช่น การใช้กฎทางภาษาศาสตร์(Linguistic rules), รูปแบบเฉพาะของการระบุชื่อ (patterns) แต่ NEXAS ไม่ได้ใช้วิธีในการสกัดโดยวิธีดังกล่าวนี้เลย เนื่องจากเหตุผลต่อไปนี้คือ 1) เว็บเพจที่หลากหลายมากเกินไปที่จะใช้รูปแบบทางภาษาศาสตร์ (linguistic pattern) ในการสกัด 2) ชื่อคนที่ปรากฏอยู่นั้นเว็บเพจนั้นไม่ได้อยู่ในประโยคที่สมบูรณ์ แต่เป็นข้อความสั้นๆประกฏอยู่ในโครงสร้างแบบ ตาราง และ สิสต์ โดยส่วนใหญ่

ดังนั้น NEXAS ใช้วิธี Dictionary-centric approach ซึ่งมีขั้นตอนดังนี้

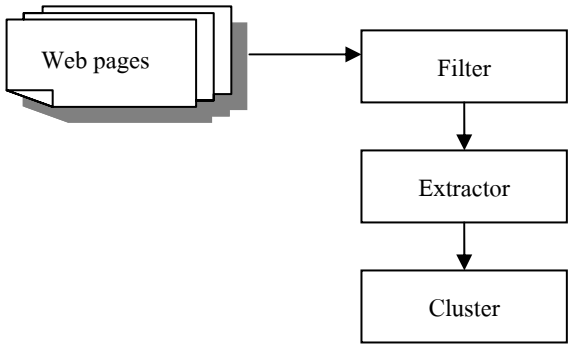
1. กำจัด HTML tag
2. ตัดคำและกำหนดค่า Part-of-Speech (POS) ให้แต่ละคำโดยใช้ตัววิเคราะห์ที่มีชื่อว่า MeCab [Data-intensive questing answering]
3. ค้นหารูปที่ปรากฏในลักษณะ ตัวแรกเป็น นามสกุล และ ตัวถัดไปเป็น ชื่อบุคคล เมื่อพบรูปแบบในลักษณะดังกล่าวแล้วจึงสกัดออกมาเป็นชื่อคน

การทดลองของระบบ NEXAS

ในการทดลอง NEXAS ใช้ ODIN ซึ่งเป็นระบบเสิร์ชเอนจิ้นแบบ full text search เพื่อค้นหาเว็บไซด์ที่มีลักษณะคล้ายกัน โดยใช้คำประเภทหัวข้อจำนวน 45 คำซึ่งได้มาจาก 3 หมวดหมู่ คือ ดนตรี, กีฬา และ เทคโนโลยีสารสนเทศ

Xiaojun Wan, Jianfeng Gao Mu Li และ Binggong Ding นำเสนองานวิจัย Person Resolution in Person Search Result: WebHawk (Wan, 2005)แหล่งข้อมูลของระบบ WebHawk เลือกมาจาก 200 อันดับแรก ในโปรแกรมสนทนาออนไลน์ (msn) หลังจากนั้นก็จะได้ เว็บเพจที่มีข้อมูลบุคคลมาแล้วนำไปสกัดเอาข้อมูลส่วนที่เป็น ชื่อและนามสกุลของบุคคล คำนำหน้าชื่อบุคคล องค์กร, อีเมลล์ และ เบอร์โทรศัพท์ออกมา

ขั้นตอนของการทำงานแสดงดังภาพที่ 15



ภาพที่ 15 แบบจำลองการทำงานของระบบ WebHawk (Wan, 2005)

การกรองหน้าเอกสารที่ไม่ต้องการออก (Filter)

ดูจากเว็บที่มีส่วนประกอบของชื่อคน Junk page ในนี้หมายเว็บที่มีชื่อคนเป็น link แต่ link นั้นอาจไม่ได้ชี้ไปยังหน้าเอกสารที่เป็นข้อมูลของบุคคลก็เป็นได้ ยกตัวอย่างเช่น ford Disney โดยจากการศึกษาของเค้าพบว่าเว็บเพจลักษณะเช่นนี้มีมากถึง 35 % ใน 100 rank แรกที่ได้มาจากการค้นหาโปรแกรมสนทนาออนไลน์ (msn) ดังนั้นปัญหา junk page จึงเป็นปัญหาใหญ่ เพราะว่ามันจะ

ก่อให้เกิดปัญหาในกระบวนการต่อไป แล้วทำให้ผู้ใช้ไม่ได้ข้อมูลที่เกี่ยวข้องกับบุคคล แต่น่าเสียดายที่ปัญหานี้ไม่ได้รับการวิเคราะห์ในงานวิจัยที่ได้สำรวจมา

โดยในงานวิจัยนี้ (Web hawk) จะมอง junk page ในรูปแบบ ของ binary classification problem และนำเทคนิค SVM มาใช้ในการแก้ปัญหานี้

การสกัดข้อมูลที่เกี่ยวข้องกับบุคคล (Extractor: Inducing Personal information)

สำหรับงานของ webhawk มีลักษณะเป็นแบบ Query oriented โดยจะสกัดเอาเฉพาะ personal profile สำหรับวิธีในการสกัดนั้นเป็นวิธีการแบบ hybrid โดยมีการผสมกันระหว่าง pattern matching และ mutual reinforcement learning โดยเค้าได้แบ่งชื่อออกเป็น 3 แบบคือ แบบที่มี 3 ฟิลด์ 2 ฟิลด์ และ 1 ฟิวล์ ซึ่งแบบ 3 ฟิวล์ จะใช้วิธีการจับคู่รูปแบบเดียวกัน (pattern matching) ในการตรวจจับชื่อบุคคล ส่วนแบบ 2 ฟิลด์ และ 1 ฟิลด์ นั้นจะใช้วิธีการ mutual reinforcement learning ในการตรวจสอบ

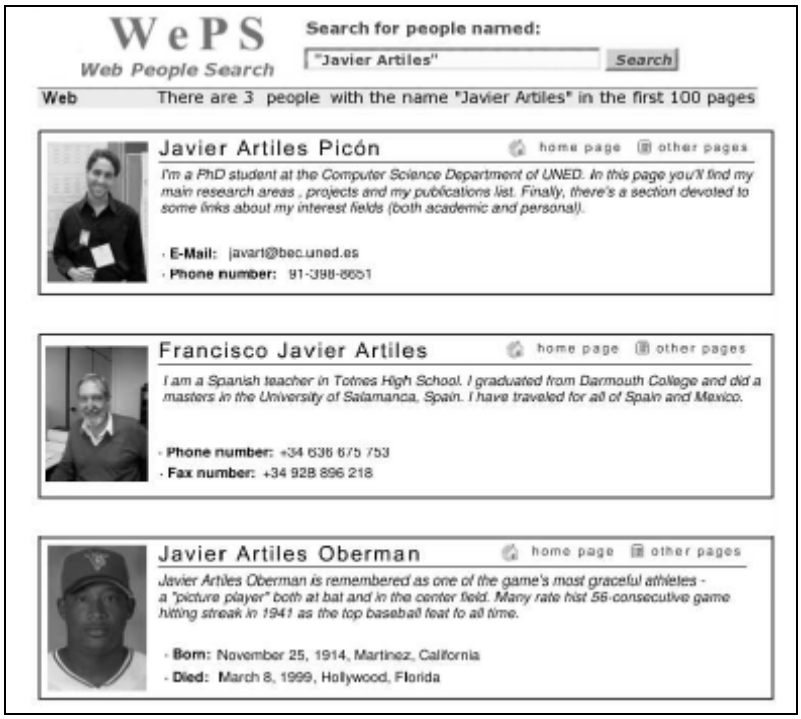
การรวมข้อมูลของบุคคลคนเดียวกันเข้าไว้ด้วยกัน (Cluster)

คลัสเตอร์เป็นวิธีการที่ใช้ในการจัดกลุ่มเอกสาร โดยข้อมูลของคนๆเดียวกันจะถูกจัดให้อยู่รวมกัน ซึ่ง webHawk ใช้วิธีอัลกอริทึมagglomerative clustering ในการจัดกลุ่มเอกสาร

ประสิทธิภาพการทำงานของ WebHawk

ในการทดลองใช้การวัดประสิทธิภาพคือ ค่าความถูกต้อง, ค่าระลึกได้ และ การวัดขนาดแบบเอฟ (F measure) ซึ่งคำนวณได้จากสมการ 2pr/(p+r) โดย p คือ ค่าระลึกได้ r คือ ค่าระลึกได้ ประสิทธิภาพในการสกัดข้อมูลของระบบ WebHawk มีดังนี้คือ การสกัดชื่อบุคคล ค่าความถูกต้อง= 93.0% ค่าระลึกได้= 92.3%, การสกัดคำนำหน้าบุคคล ค่าความถูกต้อง= 62.5% ค่าระลึกได้=55.1, การสกัดชื่อองค์กร ค่าความถูกต้อง= 22.1% ค่าระลึกได้= 33.0% การสกัดอีเมลล์แอดเดรส ค่าความถูกต้อง= 92.3% ค่าระลึกได้= 88.9% การสกัดหมายเลขโทรศัพท์ ค่าความถูกต้อง=84.3% ค่าระลึกได้= 89.6%

Javier Artiles, Julio Gonzalo และ Felisa Verdejo นำเสนองานวิจัย A Tested for People Searching Strategies in the WWW, (Artiles, 2005)ในงานวิจัยชิ้นนี้ได้ศึกษาถึงแนวทางในการวัดประสิทธิภาพของการค้นข้อมูลบุคคลที่มีอยู่ในอินเทอร์เน็ต ซึ่งสิ่งที่ผู้วิจัยมุ่งเน้นคือ การแก้ปัญหาความกำกวมของชื่อ (ชื่อเดียวกันแต่มีผู้ใช้ชื่อนั้นมากกว่า 1 คน) เนื่องมาจากผลการสำรวจที่พบว่าชื่อคนในประเทศสหรัฐอเมริกาจำนวน 90,000 ที่แตกต่างกันนั้นมีคนใช้ชื่อเหล่านั้นอยู่เป็นจำนวน 100 ล้านคน(Guha 2003) ซึ่งส่วนใหญ่แล้วผลลัพธ์จากการค้นหาบุคคลก็จะได้เว็บเพจที่มีข้อมูลของบุคคลที่ใช้ชื่อเดียวกันอยู่ ส่วนการแสดงผลลัพธ์เหล่านี้ให้แก่ผู้ใช้ โดยทั่วไประบบค้นหาข้อมูล (search engine) จะกระทำโดยการจัดลำดับของข้อมูล(ranking)แต่วิธีการที่ดีกว่านั่นคือ ระบบควรจะแสดงผลลัพธ์พร้อมกับคำบรรยายอย่างย่อเกี่ยวกับบุคคลนั้น เพราะจะทำให้ผู้ใช้สามารถเลือกข้อมูลที่ต้องการได้สะดวกกว่า ดังภาพที่ 16



ภาพที่ 16 ส่วนการแสดงข้อมูลผลลัพธ์จากการค้นคืนของระบบ WEPs (Artiles, 2005)

ผลจากงานวิจัยชิ้นนี้แบ่งออกเป็น 2 ส่วนหลักคือ การสร้างcorpus ซึ่งมีชื่อว่า WEPS (Web People Search) เพื่อใช้สำหรับการทดลอง WEPSเป็นเว็บเพจที่มีข้อมูลของคนซึ่งได้มาจากการใช้ชื่อคนเป็นคำค้นส่งไปยังระบบค้นหาข้อมูล (Search Engine) และ สร้าง baseline สำหรับการแก้ไขปัญหาความกำกวมของชื่อคน

จากการสำรวจระบบสืบค้นข้อมูลบุคคล จำนวน 6 ระบบ พบว่าระบบสืบค้นประกอบด้วยการประมวลผลย่อยทั้งหมด 3 หน่วยประมวลผล ได้แก่ การรวบรวมข้อมูลจากเว็บเพจ การสกัดข้อมูลที่เกี่ยวกับบุคคลออกจากเว็บเพจ และการจัดเก็บข้อมูลเพื่อรอการสืบค้น ในส่วนการรวบรวมข้อมูลบุคคลโดยส่วนใหญ่จะใช้เว็บครอว์เลอร์ในการรวบรวมเว็บเพจ แต่ข้อเสียประการหนึ่งของวิธีการนี้คือ ไม่สามารถกรองเอกสารตามที่ต้องการเพียงอย่างเดียวได้ ในส่วนของการสกัดข้อมูลส่วนใหญ่วิธีการที่ระบบสืบค้นข้อมูลบุคคลที่ได้ศึกษามานั้นจะใช้วิธีการสกัดเฉพาะตัว เช่น การรวบรวมข้อมูลบุคคลโดยอาศัยอีเมลล์เป็นอินพุตสำหรับสกัด ซึ่งการสกัดแบบเฉพาะเจาะจงนั้นมีข้อดีคือ สิ่งที่สกัดได้มานั้นจะมีความถูกต้องสูง เนื่องจากมีการระบุทรัพยากร (Resource) ที่แน่นอน แต่มีข้อด้อยคือ ข้อมูลที่สกัดมาได้จะเป็นข้อมูลเพียงเฉพาะบางกลุ่มไม่ครอบคลุมเอกสารที่มี ในส่วนของการจัดเก็บข้อมูล วิธีการที่เป็นที่นิยมใช้เพื่อจัดเก็บข้อมูลคือการจัดเก็บลงฐานข้อมูล ข้อดีของวิธีการนี้คือ ข้อมูลจะถูกจัดเก็บอย่างเป็นระเบียบและค้นคืนได้ในเวลารวดเร็ว แต่ข้อเสียคือ ไม่สามารถแบ่งปันข้อมูลที่ได้จัดเก็บไว้ไปสู่คอมพิวเตอร์อื่นๆได้ เนื่องจากฐานข้อมูลมีการติดตั้งแบบตายตัว และผู้ใช้จำเป็นต้องใช้คำค้นที่ตรงกับข้อมูลที่มีอยู่ในฐานข้อมูลเท่านั้นจึงจะสามารถค้นคืนข้อมูลที่ต้องการได้

สำหรับงานวิจัยนี้ใช้การรวบรวมข้อมูลโดยอาศัยเสิร์ชเอนจิ้นในการรวบรวมข้อมูล เนื่องจากช่วยลดเวลาในการคัดเลือกเว็บเพจที่มีเฉพาะข้อมูลบุคคลได้ ทำให้รวบรวมเว็บเพจได้โดยอาศัยเวลาน้อยกว่าเว็บครอว์เลอร์ ซึ่งก่อนจัดเก็บข้อมูลจะมีการตรวจสอบความคล้ายของชื่อบุคคลว่าเป็นบุคคลคนเดียวกันหรือไม่โดยการพัฒนาวิธีการตรวจสอบความคล้าย(Edit distance) ให้เหมาะสมกับการใช้งานกับภาษาไทย ส่วนสุดท้ายคือการจัดเก็บข้อมูลแบบรองรับการอนุมานและอาศัยออนโทโลจีมาช่วยในการสืบค้นข้อมูล

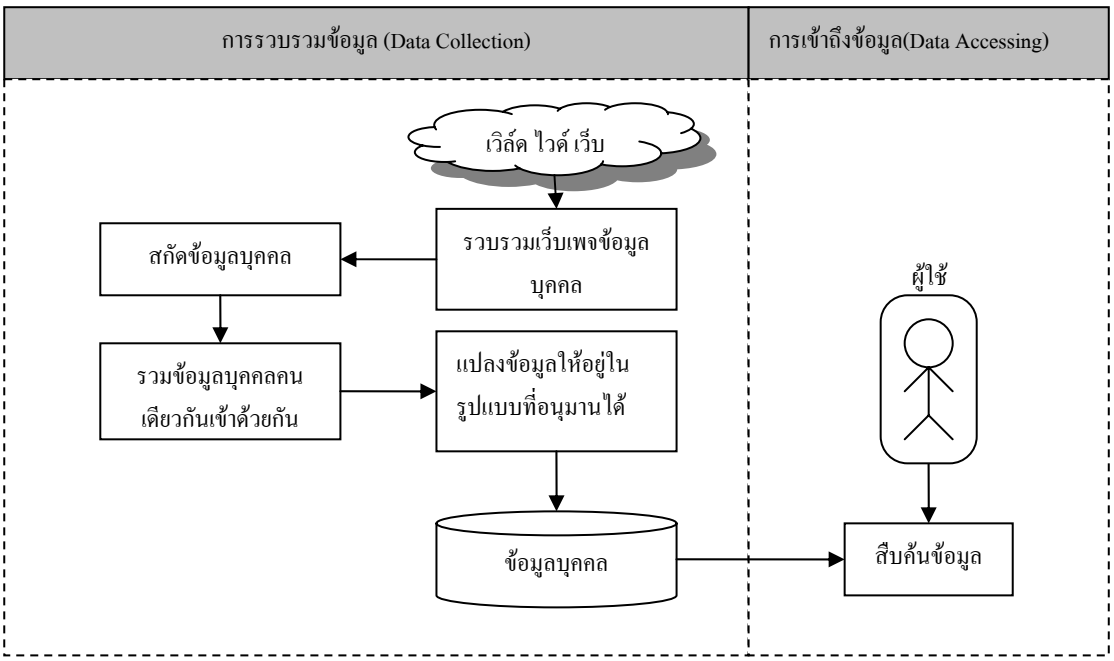
อุปกรณ์และวิธีการ

อุปกรณ์

1.	เครื่องคอมพิวเตอร์ Pentium IV 2.40 GHz หน่วยความจำ 512 MB	1 เครื่อง
2.	ระบบปฏิบัติการ Linux debian	1 ชุด
3.	ซอฟต์แวร์ Python 2.4	1 ชุด
4.	ซอฟต์แวร์ Java	1 ชุด
5.	ซอฟต์แวร์ Jena 2.3	1 ชุด

วิธีการ

ระบบสกัดและสืบค้นข้อมูลบุคคลนั้นเป็นระบบที่มีความสามารถในการรวบรวมข้อมูลของบุคคลที่มีอยู่ตามเว็บเพจต่างๆในระบบอินเทอร์เน็ตแล้วสกัดเอาข้อมูลของบุคคลออกมาแล้วจัดเก็บข้อมูลให้อยู่ในรูปแบบที่คล้ายๆกัน ดังนั้นผู้ใช้สามารถค้นหาข้อมูลบุคคลได้โดยใช้เวลาในการสืบค้นน้อยและได้รับข้อมูลที่มีเนื้อหาครบถ้วนตามความต้องการมากที่สุด อีกทั้งผู้ใช้ยังสามารถเข้าถึงข้อมูลที่ได้มาจากการอนุมานได้อีก และเพื่อให้ได้ระบบที่การทำงานได้ตามที่กล่าวมา ระบบนี้จึงแบ่งออกเป็น 2 ส่วนหลัก คือ ขั้นตอนในการรวบรวมข้อมูลบุคคลมาจัดเก็บไว้ (Data Collection) และ ส่วนของการเข้าถึงข้อมูล (Data Accessing) ดังแสดงในภาพที่ 17



ภาพที่ 17 สถาปัตยกรรมของระบบสืบค้นข้อมูลบุคคลโดยอาศัยเทคนิคอนุมาน

จากภาพที่ 17 การรวบรวมข้อมูลบุคคลเตรียมไว้เพื่อรองรับการสืบค้นจากผู้ใช้งาน ประกอบด้วย 4 ส่วนหลัก คือ การรวบรวมเว็บเพจข้อมูลบุคคล การสกัดข้อมูลบุคคล (Sirigayon, 2005) การรวมข้อมูลของบุคคลคนเดียวกันเข้าด้วยกัน การจัดเก็บข้อมูลให้อยู่ในรูปแบบที่รองรับการอนุมาน สำหรับการเข้าถึงข้อมูลนั้นประกอบด้วย ส่วนของการสืบค้นข้อมูล โดยวิธีการที่ใช้ในการสกัดข้อมูลบุคคลใช้วิธีการของ (Thamvijit,2005) สำหรับทฤษฎีที่เกี่ยวข้องได้อธิบายไว้ในบทตรวจเอกสาร งานวิจัยนี้ได้เลือกใช้เทคนิคและวิธีการที่เหมาะสมในแต่ละส่วนการทำงาน เพื่อสร้างระบบสืบค้นข้อมูลบุคคลที่มีประสิทธิภาพดี ซึ่งส่วนประกอบของระบบดังภาพที่ 17 มีรายละเอียดดังต่อไปนี้

1. รวบรวมเว็บเพจข้อมูลบุคคล

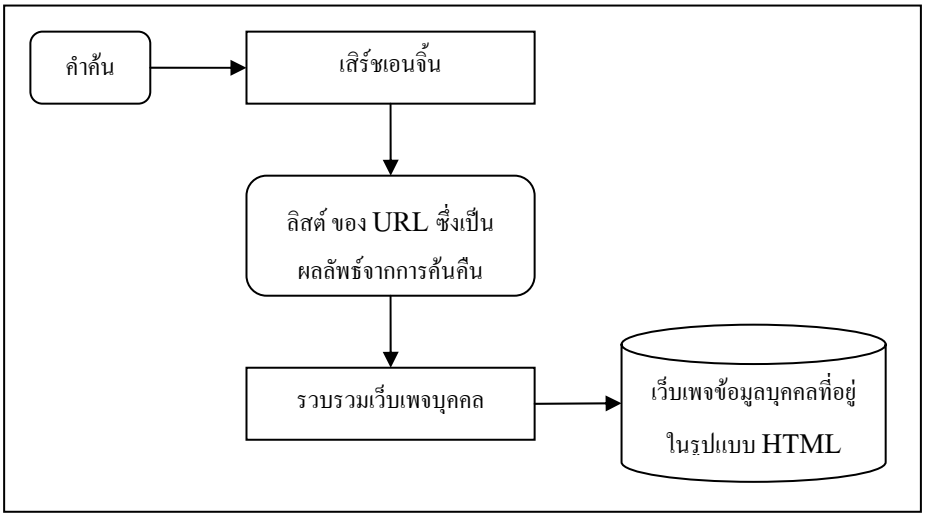
ในส่วนของการรวบรวมข้อมูลจากเว็บเพจนั้นโดยทั่วไปแล้ววิธีที่นิยมใช้กันคือ เว็บครอว์เลอร์ (Web crawler) ซึ่งหลักการทำงานของเว็บครอว์เลอร์ (web crawler) คือ ผู้ใช้จะเป็นผู้กำหนด URL เริ่มต้นเพื่อเป็นจุดเริ่มต้นให้เว็บครอว์เลอร์ดาวน์โหลดเว็บเพจ หลังจากที่เว็บครอว์เลอร์ดาวน์โหลดเว็บเพจแรกซึ่งเป็นจุดเริ่มต้นมาได้แล้ว เว็บเพจที่ถูกดาวน์โหลดมาจะถูกสกัดเอาส่วนที่ Hypertext Link ออกมาเพื่อนำไปใช้เป็น URL สำหรับการดาวน์โหลดต่อไป แต่ระบบนี้

ต้องการเฉพาะเว็บเพจที่มีข้อมูลบุคคลอยู่เท่านั้น ดังนั้นการใช้เว็บครอว์เลอร์เพื่อรวบรวมเว็บเพจจึงมีข้อเสียดังนี้คือ ต้องเสียเวลาในการคัดแยกข้อมูลที่ได้มาเพื่อเอาเฉพาะข้อมูลที่มีความเกี่ยวข้องกับบุคคลเท่านั้น

ดังนั้นจึงได้มีใช้วิธีในการรวบรวมเว็บเพจใหม่ โดยนำระบบเสิร์ชเอนจิ้นมาช่วยในการกรองข้อมูล เพื่อเลือกเอาเฉพาะเอกสารที่มีข้อมูลของบุคคลเท่านั้น ซึ่งประโยชน์ที่ได้จากวิธีการนี้คือ

1. ลดเวลาที่ต้องใช้ในการคัดแยกเว็บเพจที่มีข้อมูลบุคคลเท่านั้น
2. ลดเวลาที่ต้องใช้ในการดาวน์โหลดเว็บเพจจำนวนมากเกินความจำเป็น

โดยแบบจำลองการทำงานโดยรวมแสดงในภาพที่ 18



ภาพที่ 18 แบบจำลองการทำงานของส่วนรวบรวมข้อมูล

จากภาพที่ 18 คำค้นที่ใช้ส่งไปเสิร์ชเอนจิ้นนั้นมีรูปแบบดังต่อไปนี้ “คำนำหน้าบุคคล (เช่น นาย, นาง, นางสาว, ดร., รศ., อาจารย์) site: เว็บไซต์ขององค์กรที่ต้องการค้นหาข้อมูลบุคคล” ยกตัวอย่างเช่น “นาย OR นาง OR ดร OR อาจารย์ OR รศ site:http://www.ku.ac.th” หลังจากที่ได้ผ่านการทำงานของเสิร์ชเอนจิ้น ก็จะได้รายการของURLที่มีข้อมูลของคน URL ที่ได้มานั้นก็จะถูก

ดาวน์โหลดแล้วเก็บไว้เพื่อสกัดเอาข้อมูลที่เกี่ยวกับบุคคลออกจากเว็บเพจเหล่านี้ต่อไป สำหรับขั้นตอนในการทำงานของส่วนรวบรวมข้อมูลบุคคลมีขั้นตอนดังภาพที่ 19

```
for  title in titlelist:
        numpages = Search.gettotalpage(title)
        for page in numpage:
                urllist = getsourcepage(page)
```

ภาพที่ 19 รหัสเทียมแสดงการทำงานของโปรแกรมรวบรวมเว็บเพจบุคคล

จากภาพที่ 19 ขั้นตอนการทำงานเริ่มต้นจากการอ่านคำนำหน้าชื่อบุคคล แล้วส่งไปเป็นคำค้นให้เสิร์ชเอนจิ้นเพื่อให้เสิร์ชเอนจิ้นค้นคืน URL ที่สอดคล้องกันกับคำค้นออกมา หลังจากนั้นโปรแกรมจะคำนวณจำนวน URL ทั้งหมดที่เสิร์ชเอนจิ้นส่งคืนมา แล้วจึงสกัดเอา URL เหล่านั้นมาเก็บไว้เพื่อใช้สกัดเอาข้อมูลบุคคลออกมาในขั้นตอนต่อจากนี้

2. การสกัดข้อมูลบุคคล

การสกัดข้อมูลที่เกี่ยวข้องกับบุคคลนั้น ข้อมูลเกี่ยวกับบุคคลที่สกัดออกมานั้นมี 3 ชนิดด้วยกัน คือ ชื่อ นามสกุล, ตำแหน่ง และ องค์กรที่บุคคลผู้นั้นสังกัดอยู่ สำหรับในงานวิจัยนี้ได้นำวิธีในการสกัดข้อมูลมาจาก (Sirigayon, 2005) โดยขั้นตอนในการสกัดข้อมูลจะแสดงไว้ในภาคผนวก ก

3. การรวมข้อมูลของบุคคลคนเดียวกัน

หลังจากผ่านขั้นตอนกระบวนการสกัดข้อมูลบุคคลแล้วพบว่า ยังมีปัญหาชื่อของบุคคลคนเดียวกันแต่เขียนไม่เหมือนกันอยู่ ดังนั้นในงานวิจัยนี้ต้องการที่จะรวมข้อมูลของบุคคลคนเดียวกันโดยตรวจสอบจากความคล้ายกันของชื่อและนามสกุลของบุคคล และวิธีที่ใช้ในการตรวจสอบความคล้ายกันของสตริง (String) แต่วิธีการที่มีอยู่นั้น ยังไม่เหมาะสมแก่การตรวจสอบความคล้ายกันของอักขระภาษาไทย ดังที่ได้นำเสนอไปในส่วนของการตรวจเอกสาร โดยต้องค้นหาคำตอบก่อนว่า “มีบุคคลที่มีชื่อและนามสกุลซ้ำกันหรือไม่ เป็นจำนวนเท่าไร” ซึ่งผลของการสำรวจแสดงในตารางที่ 2

ตารางที่ 2 ผลการสำรวจบุคคลที่มีชื่อและนามสกุลซ้ำกัน

ประเภทของชื่อที่ซ้ำกัน (ชื่อ)	ผู้ชาย (%)		ผู้หญิง (%)	
	%*	**%****	**%***	**%****
2	80.65	2.58	80.68	2.20
3-5	17.06	0.5459	16.72	0.46
6-10	1.75	$55.9*10^{-3}$	1.95	$53.33*10^{-3}$
11-15	0.31	$9.92*10^{-3}$	0.37	$10.01*10^{-3}$
16-20	$10.4*10^{-2}$	$3.34*10^{-3}$	$12.06*10^{-2}$	$3.29*10^{-3}$
21-30	$5.77*10^{-2}$	$1.85*10^{-3}$	$11.25*10^{-2}$	$3.07*10^{-3}$
31-40	$3.27*10^{-2}$	$1.05*10^{-3}$	$2.57*10^{-2}$	$0.70*10^{-3}$
41-50	$1.40*10^{-2}$	$0.45*10^{-3}$	$2.57*10^{-2}$	$0.70*10^{-3}$
51-100	$1.71*10^{-2}$	$0.55*10^{-3}$	$1.13*10^{-2}$	$0.31*10^{-3}$
> 100	$0.47*10^{-2}$	$0.15*10^{-3}$	0	0

หมายเหตุ (*) คือผลการคำนวณจากชื่อที่ซ้ำกัน

(**) คือผลการคำนวณจากชื่อทั้งหมด

จากตารางที่ 2 เป็นการนำเสนอค่าสถิติของบุคคลที่มีชื่อและนามสกุลซ้ำกันโดยแหล่งข้อมูลนั้นได้มาจากรายชื่อผู้มีสิทธิ์เลือกตั้ง ซึ่งเป็นบุคคลคนละคนกัน จำนวนรายชื่อทั้งหมดที่ได้นำมาสำรวจนั้นมีจำนวนประมาณ 420,000 คน โดยแบ่งเป็นผู้ชายจำนวน 227981 คน และผู้หญิงจำนวน 202743 คน ซึ่งจากค่าสถิติในตารางที่ 2 นั้นพบว่า จำนวนชื่อของผู้ชายที่ซ้ำกัน 2 ชื่อ คิดเป็น 2.58 % ของชื่อผู้ชายทั้งหมด และมีเปอร์เซ็นต์น้อยลงเมื่อจำนวนชื่อที่ซ้ำกันมีจำนวนมากขึ้น สำหรับจำนวนชื่อของผู้หญิงที่ซ้ำกัน 2 ชื่อ คิดเป็น 2.20 % ของชื่อผู้หญิงทั้งหมด และมีเปอร์เซ็นต์น้อยลงเมื่อจำนวนชื่อที่ซ้ำกันมีจำนวนมากขึ้น จากค่าสถิติแสดงให้เห็นว่าอัตราส่วนของบุคคลที่ใช้ชื่อและนามสกุลซ้ำกันต่อจำนวนชื่อทั้งหมดนั้นมีค่าน้อยมาก จึงทำให้สรุปได้ว่าคนไทยที่ไม่ใช่บุคคลคนเดียวกันแล้วใช้ชื่อซ้ำกันนั้น พบได้น้อยมาก ดังนั้นขั้นตอนการตรวจสอบความคล้ายกันของชื่อบุคคลจึงไม่ต้องการที่จะแก้ไขปัญหานี้

ขั้นตอนการตรวจสอบความคล้ายของชื่อบุคคล

เนื่องจากวิธีการตรวจสอบความคล้ายกันของสตริง (String) แบบ edit distance ไม่มีประสิทธิภาพและความเหมาะสมในการตรวจสอบความคล้ายของสตริงในภาษาไทย ด้วยเหตุผล

ดังที่ได้กล่าวไว้ในหัวข้อการตรวจเอกสาร ดังนั้นขั้นตอนการตรวจสอบความคล้ายของชื่อบุคคลที่งานวิจัยนี้ได้พัฒนาขึ้นนั้นใช้พื้นฐานของวิธีการของ edit distance แต่ได้มีการปรับเปลี่ยนวิธีการในการคิดคะแนนโดยการเพิ่มน้ำหนักที่เหมาะสมต่อการใช้งานร่วมกันเอกสารภาษาไทย ซึ่งสำหรับขั้นตอนการทำงานของการตรวจสอบความคล้ายของชื่อบุคคลได้แสดงดังภาพที่ 20

```
Initialize:
    Pn1 = Person name & surname 1
    Pn2 = Person name & surname 2
Function:
    Precondition_filtering()
    similarity_check()
        - word_component_extractiont()
        - Similarity_calculation()
Process:
    checker = Precondition_filtering(Pn1,Pn2)
    if(checker == 'approve'):
        editdistance_value = similarity_check()
    else:
        similarity = 'mis_match'
    return editdistance_value
```

ภาพที่ 20 อัลกอริทึมจำลองการทำงานของการตรวจสอบความคล้ายของชื่อบุคคล

จากภาพที่ 20 อินพุตของระบบจะประกอบด้วยชื่อและนามสกุลของบุคคลจำนวน 2 คน ซึ่งเป็นสิ่งที่จะถูกนำมาเปรียบเทียบความคล้ายว่าเป็นชื่อของบุคคลคนเดียวกันหรือไม่ ขั้นตอนต่อไปคือ การเปรียบเทียบความคล้าย ซึ่งวิธีในการตรวจสอบความคล้ายประกอบด้วยฟังก์ชั้นการทำงานหลัก 2 ส่วนคือ การตรวจสอบความคล้ายแบบหยาบ (Precondition_checking) และการตรวจสอบความคล้ายแบบละเอียด (Similarity_checking)

การตรวจสอบความคล้ายแบบหยาบ (Precondition_checking)

ขั้นตอนการทำงานในส่วนนี้คือการคัดแยกชื่อบุคคลที่มีความแตกต่างกันมากเกินไปออก เพื่อไม่ให้ชื่อบุคคลต้องถูกส่งไปคำนวณหาค่าความคล้ายแบบละเอียด ซึ่งเป็นขั้นตอนที่ใช้เวลาในการตรวจสอบนานกว่า สิ่งที่ขั้นตอนนี้ทำการตรวจสอบคือ ความแตกต่างระหว่างความยาวของสตริง(ในที่นี้คือชื่อและนามสกุล) และ จำนวนอักขระที่ซ้ำกันระหว่างสตริง ดังสมการในภาพที่ 21

Let b = base word

p = compare word

$C_{b(i)}$ = i^{th} character of base word

$C_{p(j)}$ = j^{th} character of compare word

L(W) = word length, i.e. character count

p' = number of characters in p that are common with b

สมการ (1) $L(b)-1/2(\max(L(b),L(p))) < L(p) < L(b)+1/2(\max(L(b),L(p))$

สมการ (2) $p' >= 0.5*(\min(L(b),L(p))$

เงื่อนไข 1 : $|L(a)-L(b)| < 0.5 \times \max(L(a),L(b))$

เงื่อนไข 2 : $p' >= 0.5 \times \min(L(a),L(b))$

ภาพที่ 21 สมการแสดงขั้นตอนการตรวจสอบความคล้ายแบบหยาบ

สมการที่ 1 จากภาพที่ 21 ใช้เพื่อตรวจสอบความยาวของสตริงว่าแตกต่างกันมากเกินไปหรือไม่ และในสมการที่ 2 ใช้เพื่อตรวจสอบจำนวนของตัวอักษรที่ซ้ำกัน ซึ่งสตริงผ่านการตรวจสอบความคล้ายในสมการที่ 2 คือสตริงที่มีจำนวนตัวอักษรซ้ำกันมากตามเงื่อนไขที่ 2

การตรวจสอบความคล้ายแบบละเอียด(Similarity_check)

ในขั้นตอนนี้เป็นการตรวจสอบความคล้ายของชื่อบุคคลแบบละเอียด โดยในงานวิจัยนี้ได้ใช้นำเสนอวิธีการตรวจสอบความคล้ายของสตริงโดยพัฒนาวิธีการ Edit-distance ซึ่งเป็นวิธีในการตรวจสอบความคล้ายกันของสตริงแบบทั่วไป ให้มีความเหมาะสมแก่การใช้งานภาษาไทย (ดูคำอธิบายเกี่ยวกับวิธีการตรวจสอบความคล้ายกันของสตริงแบบ Edit distance ในหัวข้อการตรวจเอกสาร)

สิ่งที่ได้พัฒนาเพิ่มขึ้นมาคือวิธีในการในการคำนวณค่าคะแนนซึ่งเป็นตัวบ่งบอกความคล้าย โดยปรกติแล้วการคิดคะแนนของวิธีการ Edit distance นั้นจะคำนวณคะแนนจากการเปลี่ยนแปลงของอักขระทั้ง 3 แบบ คือ อักษรเกิน (insertion), อักษรขาด (deletion), และ อักษรแทนที่ (substitution) ซึ่งการคิดคะแนนจะคิดคะแนนเท่ากันทั้ง 3 แบบของการเปลี่ยนแปลง ดังนั้นจึงได้พัฒนาวิธีคำนวณค่าน้ำหนักในการเปลี่ยนแปลงแต่ละครั้ง ซึ่งวิธีในการคิดคำนวณน้ำหนักมีดังต่อไปนี้

<u>สิ่งที่นำมาพิจารณาเพื่อการคำนวณค่าน้ำหนัก</u>

คำในภาษาไทยสามารถมีได้มากกว่า 1 พยางค์ ซึ่งในแต่ละพยางค์นั้นส่วนใหญ่แล้วจะมีองค์ประกอบ 5 ส่วนด้วยกัน (แต่ไม่จำเป็นต้องมี 5 องค์ประกอบพร้อมกัน) ได้แก่ พยัญชนะต้น, สระ, ตัวสะกด, วรรณยุกต์ และ การันต์ ซึ่งตัวอักษรในภาษาไทยจะมีเสียงที่แตกต่างกันเมื่อทำหน้าที่แตกต่างกัน ตัวอย่างเช่น อักษร “น” กับ “ล” ดังแสดงในตารางที่ 3

<u>ตารางที่ 3</u> ตัวอย่างเสียงของอักษรเมื่อทำหน้าที่แตกต่างกัน

น and ล เมื่อทำหน้าที่เป็นตัวสะกด	กน	น เสียง = n
	กล	ล เสียง = n
น and ล เมื่อทำหน้าที่เป็นพยัญชนะต้น	นก	น เสียง = n
	ลก	ล เสียง = l

ในการกำหนดค่าน้ำหนักที่เหมาะสมแก่การเปลี่ยนแปลงในแต่ละแบบจึงจำเป็นต้องตรวจสอบก่อนว่าอักษรแต่ละตัวในพยางค์นั้นมีหน้าที่เป็นอะไร เมื่อทราบแล้วว่าอักษรแต่ละตัวทำหน้าที่เป็นอะไร (พยัญชนะต้น, สระ, ตัวสะกด, วรรณยุกต์, และการันต์) ก็จะใช้ข้อมูลกลุ่มของตัวอักษรที่มีเสียงคล้ายกัน โดยแบ่งตามหน้าที่ของตัวอักษร ดังแสดงในตารางที่ 4, ตารางที่ 5 และตารางที่ 6 เพื่อคำนวณค่าน้ำหนักของการเปลี่ยนแปลงต่อไป

ตารางที่ 4 กลุ่มของตัวอักษรที่เสียงคล้ายกันเมื่อทำหน้าที่เป็นพยัญชนะต้น

ลำดับที่	อักษร	ลำดับที่	อักษร	ลำดับที่	อักษร	ลำดับที่	อักษร
1	ก	8	ซ	15	บ	22	ร
2	ข ฃ	9	ญ ย	16	ป	23	ล ฬ
3	ค ฅ ฆ	10	ฎ ด ฑ	17	ผ	24	ว
4	ง	11	ฏ ต	18	ฝ	25	ษ ศ ส
5	จ	12	ฐ ถ	19	พ ภ	26	ห
6	ฉ	13	ท ธ ฑ ฒ	20	ฟ	27	อ
7	ช ฌ	14	ณ น	21	ม	28	ฮ

ตารางที่ 5 กลุ่มของตัวอักษรที่เสียงคล้ายกันเมื่อทำหน้าที่เป็นตัวสะกด

ลำดับที่	อักษร	ลำดับที่	อักษร
1	ก ข ฃ ค ฅ ฆ	5	บ ป ฝ พ ฟ ภ
2	ง	6	ม
3	จ ฉ ช ซ ฌ ฎ ฏ ฐ ฑ ฒ ด ต ถ ท ธ ศ ษ ส	7	ย
4	ญ ณ น ร ล ฬ	8	ว

ตารางที่ 6 กลุ่มของสระที่มีเสียงคล้ายกัน

ลำดับที่	สระ
1	◌ั ย ใ ไ
2	◌ั ม ◌ำ
3	-รร ◌ั or [◌ั +any final consonant in group 4]

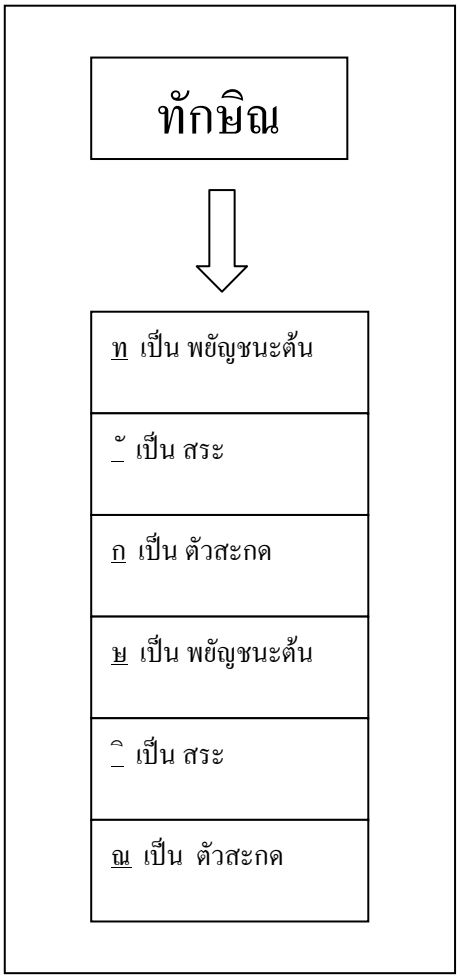
สำหรับ วรรณยุกต์ ได้แก่ : “ ◌่ ”, “ ◌้ ”, “ ◌๊ ”, “ ◌๋ ” ของชื่อบุคคลภาษาไทยแล้ว พบว่าความสับสนในการใช้มีน้อยมากจึงไม่นำมาคิดค่าน้ำหนักในการคำนวณค่าความความคล้าย แต่ตัวการันต์เป็นสิ่งที่มักจะใช้สับสนอยู่บ่อย ดังนั้นการเปลี่ยนแปลงที่เกิดขึ้นกันตัวการันต์จึงไม่ถูกนำมาคำนวณค่าน้ำหนัก ตัวอย่างเช่น “สิง” กับ “สิงห์” ค่าคะแนนที่เกิดจากการคำนวณจะออกมาเท่ากัน

ขั้นตอนในการตรวจสอบความคล้ายของชื่ออย่างละเอียด

ประกอบด้วยขั้นตอนย่อย 2 ส่วนคือ

การระบุหน้าที่ให้แก่อักษรแต่ละตัว (Word component extraction)

จากที่ได้กล่าวมาก่อนหน้านี้ว่าวิธีที่ใช้ในการคำนวณค่าน้ำหนักนั้นพิจารณาจากกลุ่มของตัวอักษรที่มีเสียงคล้ายกัน โดยแบ่งแยกตามหน้าที่ ดังนั้นจึงจำเป็นต้องรู้ก่อนว่าอักษรนั้นทำหน้าที่เป็นอะไร ในขั้นตอนนี้จึงเป็นขั้นตอนระบุหน้าที่ให้แต่ละอักษร ดังตัวอย่างในภาพที่ 22



ภาพที่ 22 ตัวอย่างการกำหนดหน้าที่ให้แต่ละอักษรในภาษาไทย

การคำนวณหาค่าความคล้าย (Similarity calculation)

ขั้นตอนในการตรวจความคล้ายกันของชื่อคนอย่างละเอียดนั้น ระบบนี้ได้นำวิธีการ edit distance มาประยุกต์ใช้งาน โดยมีการปรับเปลี่ยนวิธีในการคำนวณคะแนนความคล้าย โดยกำหนดให้ความแตกต่างของอักขระที่เกิดขึ้นในแต่ละครั้งมีคะแนนไม่เท่ากัน ซึ่งขึ้นอยู่กับคุณสมบัติของแต่ละอักขระ โดยในภาพที่ 23 จะแสดงสมการที่ใช้ในการคำนวณคะแนน โดยสมการที่ 4 เป็นแบบของการพิมพ์เกินและพิมพ์ตก สมการที่ 3 เป็นแบบของการพิมพ์ผิด

กำหนดให้ Fn.close_form(x,y) ใช้เพื่อตรวจสอบว่าอักษร x และ y มีความคล้ายกันหรือไม่

Fn. in_group(x,y) ใช้เพื่อตรวจสอบว่าอักษร x และ y อยู่ในกลุ่มเสียงเดียวกันหรือไม่

0 : if x = y

สมการ (3) subs_cost(x,y) = 1 : if in_group(x,y) = true

2 : if close_form(x,y) = true

3 : else

สมการ (4) insert_cost(x,y) or delete_cost(x,y) = 0 ; if x = y

1 ; if in_group(x,y) = true

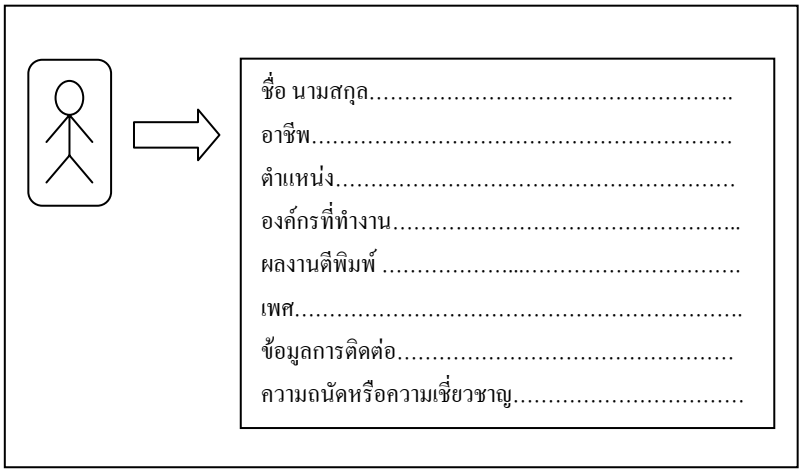
2 ; else

ภาพที่ 23 สมการแสดงขั้นตอนการตรวจสอบความคล้ายอย่างละเอียด

4. การจัดเก็บข้อมูลบุคคลเพื่อรองรับการอนุมาน

ในองค์กรหนึ่งย่อมประกอบด้วยบุคคลซึ่งทำงานอยู่ในองค์กรนั้นๆ โดยแต่ละคนย่อมมีหน้าที่รับผิดชอบต่างกัน ซึ่งอาจเหมือนหรือแตกต่างกันก็ได้ ถ้าหากมีบุคคลภายนอกต้องการจะค้นหาข้อมูลบุคคลภายในองค์กรแล้ว สิ่งที่ต้องการมากที่สุดคือ ความสามารถหรือความถนัดของผู้ที่ต้องการค้นหา เช่น “ต้องการค้นหาบุคคลผู้ซึ่งสามารถให้คำปรึกษาเกี่ยวกับการปลูกผักปลอดสารพิษได้” นอกจากนี้ยังมีตัวอย่างที่แสดงให้เห็นว่า ความสามารถของบุคคลเป็นสิ่งที่บุคคลอื่นอยากรู้มากคือ การประกาศรับสมัครงาน เช่น “มีความสามารถในการเขียนโปรแกรมภาษาซีได้”

ข้อมูลบุคคลเป็นข้อมูลที่ใช้บรรยายว่าแท้จริงแล้วบุคคลนั้นคือใคร มีความสามารถอะไรบ้าง ดังตัวอย่างในภาพที่ 24



ภาพที่ 24 ตัวอย่างของข้อมูลเกี่ยวกับบุคคล

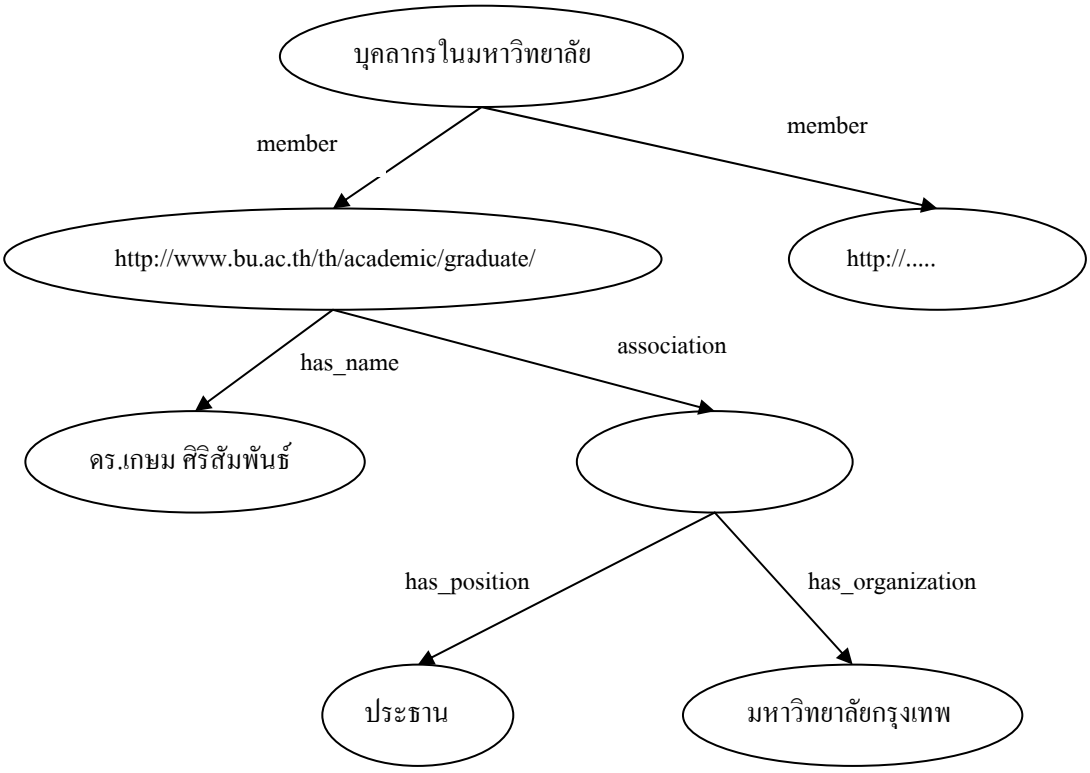
จากภาพที่ 24 เป็นตัวอย่างข้อมูลที่ใช้บรรยายถึงบุคคล จากการสำรวจพบว่า ข้อมูลที่มีอยู่ในเว็บเพจนั้นใช้คำในการบรรยายข้อมูลของตัวเองไม่ตรงกับข้อมูลที่ผู้อื่นต้องการค้นหา ตัวอย่างเช่น ในช่วงที่มีการระบาดของไข้หวัดนก มีผู้ใช้ต้องการค้นหาผู้เชี่ยวชาญในการแก้ปัญหาไข้หวัดนก ในขณะที่ในองค์กรหนึ่งระบุว่ามีบุคคลที่มีความเชี่ยวชาญเกี่ยวกับไวรัส H5N1 ดังนั้นผู้ใช้จึงไม่สามารถค้นหาผู้เชี่ยวชาญผู้นี้ได้ ทั้งที่ไวรัส H5N1มีความเกี่ยวข้องกันอย่างมากโดย H5N1 เป็นไวรัสที่ทำให้เกิดโรคไข้หวัดนก ดังนั้นในงานวิจัยนี้มุ่งเน้นที่จะแก้ไขปัญหาการค้นหาบุคคลแต่ผู้ใช้ เลือกใช้คำที่ใช้ในการค้นหาไม่ตรงกันกับที่บุคคลผู้นั้นได้เขียนบรรยายตัวเองไว้

จากข้อดีของการจัดเก็บข้อมูลแบบ OWL (web ontology language) ซึ่งเป็นจัดเก็บข้อมูลที่รองรับการอนุมาน และออนโทโลจี (ontology) ตามที่ได้บรรยายไว้ในส่วนของการการตรวจเอกสาร งานวิจัยนี้จึงนำวิธีการดังกล่าวมาใช้ในการจัดเก็บข้อมูลบุคคล และได้ออกแบบกฎที่ใช้สำหรับการอนุมาน ทำให้ผู้ใช้สามารถค้นหาบุคคลได้ตรงกับความต้องการมากยิ่งขึ้น

ในงานวิจัยนี้ได้ใช้กรอบในการทำงานของ JENA (Java framework for building Semantic Web applications) ซึ่งพัฒนาโดยบริษัท HP (www.hp.com) เพื่อความสะดวกในการจัดการข้อมูลที่ถูกจัดเก็บในรูปแบบ RDF, OWL และการอนุมาน

โมเดลในการจัดเก็บข้อมูล

ในภาษาอาร์ดีเอฟ (RDF) การจัดเก็บข้อมูลจะสามารถเขียนให้อยู่ในรูปแบบของโมเดลได้ โดยลักษณะของโมเดลจะอยู่ในของกราฟซึ่งประกอบด้วยโหนดกับเอจ ในงานวิจัยนี้ได้ออกแบบโมเดลของข้อมูลที่ต้องการจัดเก็บในโครงสร้างแบบอาร์ดีเอฟ ดังแสดงในภาพที่ 25



ภาพที่ 25 โมเดลของข้อมูลบุคคลที่ต้องการจัดเก็บในโครงสร้างอาร์ดีเอฟ

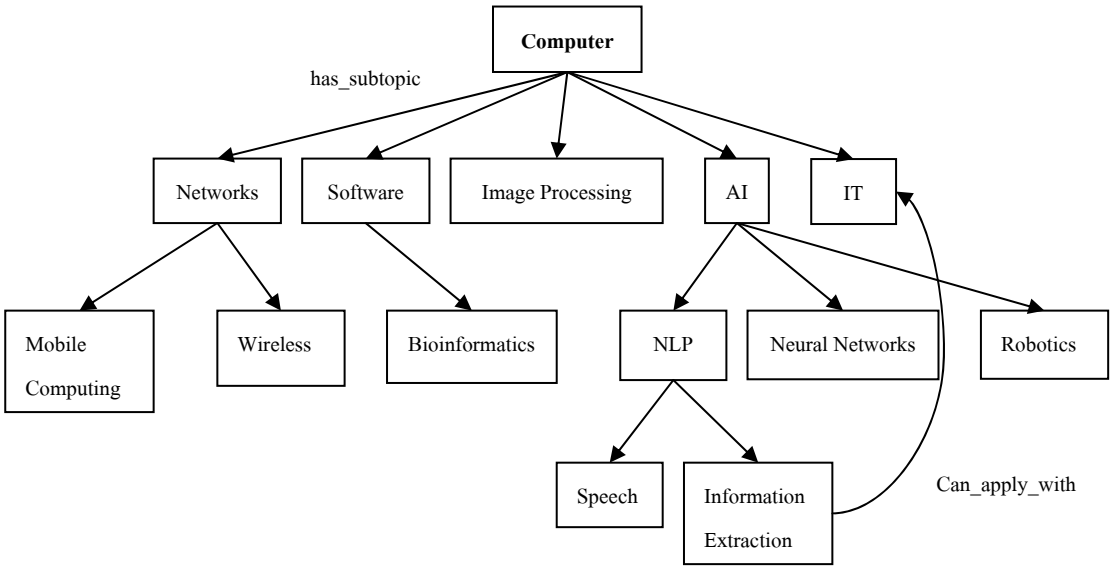
จากภาพที่ 25 แสดงให้เห็นถึงโมเดลในการจัดเก็บข้อมูลของคนหนึ่งคนในรูปแบบของอาร์ดีเอฟ โดยความสัมพันธ์ที่ได้กำหนดไว้เพื่อใช้ในการบรรยายข้อมูลบุคคลได้แก่ member, has_name, association, has_position, has_organization และจากแบบจำลองนี้สามารถเขียนให้อยู่ในรูปแบบวากยสัมพันธ์ของภาษาอาร์ดีเอฟได้ดังภาพที่ 26

```
<rdf:RDF
   xmlns:j.0="http://naist.cpe.ku.ac.th/"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
  <rdf:Description rdf:nodeID="A0">
   <j.0:has_posit>ประธานสภาวิชาการ</j.0:has_posit>
  </rdf:Description>
  <rdf:Description rdf:nodeID="A1">
   <j.0:has_organ>สภาฯ</j.0:has_organ>
   <j.0:has_posit>ตำแหน่งปัจจุบัน</j.0:has_posit>
  </rdf:Description>
  <rdf:Description rdf:nodeID="A2">
   <j.0:has_posit>บัณฑิตศึกษา</j.0:has_posit>
  </rdf:Description>
  <rdf:Description rdf:nodeID="A3">
   <j.0:has_posit>ประธานกรรมการประจำสาขาวิชาเศรษฐศาสตร์</j.0:has_posit>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.mju.ac.th/about/Admin-Staff/Admin-thai.htm#ผศ.ดร.พิสุทธิ์ เนียมทรัพย์">
   <j.0:has_name>ผศ.ดร.พิสุทธิ์ เนียมทรัพย์</j.0:has_name>
   <j.0:has_association rdf:nodeID="A4"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.mju.ac.th/about/admin-sapa/#ดร.กฤษณพงศ์ กีรติกร">
   <j.0:has_name>ดร.กฤษณพงศ์ กีรติกร</j.0:has_name>
   <j.0:has_association rdf:nodeID="A5"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.payap.ac.th/news/alumni/welcome.htm#ดร.บุญทอง ภู่เจริญ อธิการบดี">
   <j.0:has_name>ดร.บุญทอง ภู่เจริญ อธิการบดี</j.0:has_name>
   <j.0:has_association rdf:nodeID="A6"/>
  </rdf:Description>
  <rdf:Description
```

ภาพที่ 26 ข้อมูลบุคคลที่ถูกจัดเก็บในภาษาอาร์ดีเอฟ

จากภาพที่ 26 แสดงให้เห็นถึงข้อมูลบุคคลที่ถูกจัดเก็บอยู่ในรูปแบบของอาร์ดีเอฟ ซึ่งที่แสดงในภาพที่ 26 นี้เกิดขึ้นจากการประมวลผลโดยใช้กรอบระบบของเจน่า (Jena) ซึ่งถูกออกแบบมาเพื่อให้รองรับการจัดการบนข้อมูลที่ถูกจัดเก็บในภาษาอาร์ดีเอฟและภาษาอาวล์

หลังจากที่ได้จัดข้อมูลบุคคลตามโครงสร้างดังกล่าวเป็นที่เรียบร้อยแล้ว ข้อมูลที่ถูกจัดเก็บนี้จะมีการทำงานเชื่อมต่อกับออนโทโลจี เพื่อเป็นช่องทางในการเข้าถึงข้อมูลของผู้ใช้ให้มากยิ่งขึ้น ยกตัวอย่างเช่น ผู้ใช้ต้องการค้นหาบุคคลที่มีความเชี่ยวชาญเกี่ยวกับเทคโนโลยีเครือข่ายไร้สาย แต่ภายในข้อมูลจริงที่ได้จัดเก็บนั้นมีเพียงบุคคลที่มีความสนใจในเทคโนโลยีเครือข่ายเท่านั้น แต่เป็นที่ทราบกันดีว่าเทคโนโลยีเครือข่ายไร้สายเป็นส่วนหนึ่งในเทคโนโลยีเครือข่าย ดังนั้น บุคคลที่มีความสนใจในเทคโนโลยีเครือข่ายจึงสามารถเป็นคำตอบในคำถามนี้ได้ ซึ่งการจะทำให้เกิดการเชื่อมโยงในลักษณะนี้สามารถกระทำได้โดยอาศัยออนโทโลจีของสาขาวิชาย่อยของคอมพิวเตอร์ ดังแสดงในภาพที่ 27



ภาพที่ 27 ออนโทโลจีด้านคอมพิวเตอร์

จากภาพที่ 27 เป็นแบบจำลองของออนโทโลจีของคอมพิวเตอร์ ซึ่งออนโทโลจีนั้นมีส่วนประกอบ 2 อย่างคือ พจน์ (Term) และความสัมพันธ์ (Relation) ระหว่างพจน์ จากตัวอย่างของออนโทโลจีในภาพที่ 27 ความสัมพันธ์ระหว่างพจน์นั้นคือ has_subtopic และ can_apply_with ซึ่งจะเห็นความสัมพันธ์ระหว่างพจน์นั้นมีได้มากกว่า 1

เมื่อนำออนโทโลจีนี้ใช้งานร่วมระบบสืบค้นข้อมูลบุคคลแล้ว ผู้ใช้ไม่จำเป็นต้องใช้คีย์เวิร์ดตรงตามข้อมูลจริงที่มีอยู่ โดยออนโทโลจีจะทำหน้าที่ขยายคำค้นที่ใกล้เคียงกับคำค้นที่ผู้ใช้เลือกใช้

มากที่สุด เพื่อให้ผู้ใช้ได้รับข้อมูลที่เกี่ยวข้องแทนที่จะไม่ได้รับข้อมูลใดๆเลยจากการสืบค้น ในกรณีที่ผู้ใช้เลือกใช้คำค้นไม่ตรงกับข้อมูลที่จัดเก็บไว้

ออนโทโลจีที่อยู่ในรูปแบบที่จะนำไปใช้งานจริงนั้น ในงานวิจัยนี้ได้เลือกใช้ภาษาอาวล์ (OWL) ในการจัดเก็บออนโทโลจี ซึ่งมีลักษณะดังภาพที่ 28

```
<?xml version="1.0"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:owl="http://www.w3.org/2002/07/owl#"
   xmlns="http://www.owl-ontologies.com/unnamed.owl#"
 xml:base="http://www.owl-ontologies.com/unnamed.owl">
 <owl:Ontology rdf:about=""/>
 <owl:Class rdf:ID="Bioinformatic">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="Software"/>
  </rdfs:subClassOf>
 </owl:Class>
 <owl:Class rdf:ID="Mobile_computing">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="Network"/>
  </rdfs:subClassOf>
 </owl:Class>
 <owl:Class rdf:about="#Software">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="computer"/>
  </rdfs:subClassOf>
 </owl:Class>
 <owl:Class rdf:about="#Network">
  <rdfs:subClassOf rdf:resource="#computer"/>
 </owl:Class>
 <owl:Class rdf:ID="Neural_network">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="AI"/>
  </rdfs:subClassOf>
 </owl:Class>
 <owl:Class rdf:ID="Wireless">
  <rdfs:subClassOf rdf:resource="#Network"/>
```

```
 </owl:Class>
 <owl:Class rdf:ID="IT">
  <rdfs:subClassOf rdf:resource="#computer"/>
 </owl:Class>
 <owl:Class rdf:ID="Robotic">
  <rdfs:subClassOf>
    <owl:Class rdf:about="#AI"/>
  </rdfs:subClassOf>
 </owl:Class>
 <owl:Class rdf:ID="NLP">
  <rdfs:subClassOf>
    <owl:Class rdf:about="#AI"/>
  </rdfs:subClassOf>
 </owl:Class>
 <owl:Class rdf:ID="Image_Processing">
  <rdfs:subClassOf rdf:resource="#computer"/>
 </owl:Class>
 <owl:Class rdf:about="#AI">
  <rdfs:subClassOf rdf:resource="#computer"/>
 </owl:Class>
 <owl:Class rdf:ID="Speech">
  <rdfs:subClassOf rdf:resource="#NLP"/>
 </owl:Class>
 <owl:Class rdf:ID="Information_extraction">
  <rdfs:subClassOf rdf:resource="#NLP"/>
 </owl:Class>
</rdf:RDF>
```

ภาพที่ 28 ออนโทโลจีด้านคอมพิวเตอร์ในรูปแบบของภาษาอาวล์

จากภาพที่ 28 เป็นออนโทโลจีของสาขาความเชี่ยวชาญด้านคอมพิวเตอร์ซึ่งถูกจัดเก็บในภาษาอาวล์ ซึ่งในการนำออนโทโลจีไปใช้งานในระบบสืบค้นข้อมูลบุคคลของงานวิจัยนี้นั้น จะช่วยให้ผู้ใช้ได้รับข้อมูลบุคคลที่เกี่ยวข้องโดยไม่จำเป็นต้องใช้คำค้นให้ตรงกันกับข้อมูลจริงที่มีอยู่ ตัวอย่างเช่น ผู้ใช้ต้องการค้นหาผู้เชี่ยวชาญเกี่ยวกับหุ่นยนต์ แต่ในข้อมูลจริงไม่มีข้อมูลของหุ่นยนต์อยู่เลย ออนโทโลจีจะทำหน้าที่ในการช่วยขยายความหมายของหุ่นยนต์ ว่าหุ่นยนต์เกี่ยวข้องกับทฤษฎีปัญญาประดิษฐ์ และ เกี่ยวข้องกับคอมพิวเตอร์ฮาร์ดแวร์ ดังนั้นแทนที่ผู้ใช้จะไม่รับข้อมูล

ใดๆจากการสืบค้นเลย ผู้ใช้จะได้รับข้อมูลของผู้เชี่ยวชาญด้านปัญญาประดิษฐ์และคอมพิวเตอร์ฮาร์ดแวร์แทน

นอกจากการนำออนโทโลจีมาช่วยเพิ่มช่องทางในการสืบค้นให้แก่ผู้ใช้แล้ว ในงานวิจัยนี้ยังใช้การอนุมานมาช่วยให้ข้อมูลที่ได้จัดเก็บไว้เพิ่มขึ้น ซึ่งส่วนที่เพิ่มขึ้นนั้นเป็นข้อมูลที่เกิดจากการอนุมาน โดยการอนุมานที่ใช้ในงานวิจัยนี้จะเป็นการอนุมานโดยอาศัยกฎ (Rules based) ซึ่งจะทำให้ผู้ใช้สามารถได้รับข้อมูลบางอย่างที่ซ่อนอยู่จากข้อมูลที่มีอยู่จริง ในงานวิจัยนี้ได้กฎในการอนุมานดังตารางที่ 7

ตารางที่ 7 กฎที่ใช้สำหรับการอนุมานในระบบสืบค้นข้อมูลบุคคล

ชื่อ	กฎสำหรับการอนุมาน
กฎการถ่ายทอด	(?:A ?:p ?B), (?:B ?:p ?:C) → (?:A ?:p ?:C)
กฎการสมมาตร	(?:Y ?:p ?:X) → (?:X ?:p ?:Y)
กฎการหาผู้เชี่ยวชาญ1	(?:A ?:has_paper ?:B),(?:B ?:has_topic ?:C) → (?:A ?:expert_in ?:C)
กฎการหาผู้เชี่ยวชาญ2	(?:A ?:has_title ?:ดร.),(?:A ?:has_research_topic ?:C) → (?:A ?:expert_in ?:C)
กฎการหาผู้เชี่ยวชาญ3	(?:A ?:has_title ?:รศ.),(?:A ?:has_research_topic ?:C) → (?:A ?:expert_in ?:C)
กฎการหาผู้เชี่ยวชาญ4	(?:A ?:has_title ?:รศ.ดร.),(?:A ?:has_research_topic ?:C) → (?:A ?:expert_in ?:C)

จากกฎการอนุมานจากตารางที่ 7 กฎการถ่ายทอด จะถูกใช้เมื่อความสัมพันธ์บางอย่างในข้อมูลบุคคลนั้นสามารถถ่ายทอดได้ เช่น “นายเอตำแหน่งสูงกว่านายบี และ นายบีมีตำแหน่งสูงกว่านายซี แล้ว นายเอจะมีตำแหน่งสูงกว่านายซีด้วย” กฎการสมมาตร ซึ่งถูกใช้เมื่อความสัมพันธ์ของข้อมูลบุคคลสมมาตรกัน เช่น “นายเออยู่หน่วยงานเดียวกับนายบี นายบีก็อยู่หน่วยงานเดียวกับนายเอ เช่นกัน” สุดท้ายเป็นกฎการหาผู้เชี่ยวชาญ ซึ่งในงานวิจัยนี้ ได้ใช้วิธีระบุตัวผู้เชี่ยวชาญโดยดูจากผลงานตีพิมพ์ ซึ่งหมายความว่าผู้ที่มีผลงานตีพิมพ์ในหัวข้อใดๆแล้วย่อมเป็นผู้เชี่ยวชาญในหัวข้อนั้นด้วย และอีกวิธีการหนึ่งในการระบุตัวผู้เชี่ยวชาญคือ การระบุจากคำนำหน้าชื่อ เช่น ดร. รศ. รศ. ดร. โดยถ้าบุคคลใดมีคำนำหน้าชื่อเหล่านี้แล้ว และบุคคลผู้นั้นทำงานวิจัยในหัวข้อใดๆแล้ว ถือว่าบุคคลผู้นั้นเป็นผู้เชี่ยวชาญในหัวข้อที่ทำวิจัยนั้นด้วย

ผลการทดลอง

งานวิจัยนี้ได้ทำการทดลองกับเว็บเพจในหมวดหมู่ของมหาวิทยาลัย เพื่อสกัดเอาข้อมูลบุคคลที่อยู่ภายในองค์กรมหาวิทยาลัย ซึ่งสิ่งที่จะทำการสกัดคือ ชื่อ นามสกุล, ตำแหน่ง และ องค์กร หลังจากนั้นจะนำข้อมูลที่ได้มาจากการสกัดไปจัดเก็บในรูปแบบที่รองรับการอนุมาน ซึ่งมีจำนวนทั้งสิ้น 33 มหาวิทยาลัย ดังต่อไปนี้

ตารางที่ 8 รายชื่อมหาวิทยาลัยที่เป็นแหล่งข้อมูลที่ใช้ในการทดลอง

ลำดับ	รายชื่อมหาวิทยาลัย	URL
1	มหาวิทยาลัยเกษตรศาสตร์	http://www.ku.ac.th
2	มหาวิทยาลัยกรุงเทพ	http://www.bu.ac.th
3	มหาวิทยาลัยเกริก	http://www.krirk.ac.th
4	มหาวิทยาลัยเกษมบัณฑิต	http://kasem.kbu.ac.th
5	มหาวิทยาลัยขอนแก่น	http://www.kku.ac.th
6	จุฬาลงกรณ์มหาวิทยาลัย	http://www.chula.ac.th
7	มหาวิทยาลัยชินวัตร	http://www.shinawatra.ac.th
8	มหาวิทยาลัยเชียงใหม่	http://www.chiangmai.ac.th
9	มหาวิทยาลัยเซนต์จอห์น	http://www.stjohn.ac.th
10	มหาวิทยาลัยทักษิณ	http://www.tsu.ac.th
11	มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี	http://www.kmitt.ac.th
12	มหาวิทยาลัยเทคโนโลยีมหานคร	http://www.mut.ac.th
13	มหาวิทยาลัยเทคโนโลยีสุรนารี	http://www.sut.ac.th
14	มหาวิทยาลัยธรรมศาสตร์	http://www.tu.ac.th
15	มหาวิทยาลัยธุรกิจบัณฑิตย์	http://www.dpu.ac.th
16	มหาวิทยาลัยนเรศวร	http://www.nu.ac.th
17	มหาวิทยาลัยบูรพา	http://www.buu.ac.th
18	มหาวิทยาลัยพายัพ	http://www.payap.ac.th
19	มหาวิทยาลัยมหาสารคาม	http://www.msu.ac.th
20	มหาวิทยาลัยมหิดล	http://www.mahidol.ac.th
21	มหาวิทยาลัยแม่โจ้	http://www.mju.ac.th
22	มหาวิทยาลัยรังสิต	http://www.rsu.ac.th
23	มหาวิทยาลัยรามคำแหง	http://www.ru.ac.th
24	มหาวิทยาลัยวิทยาศาสตร์และเทคโนโลยีแห่งเอเชีย	http://www.ait.ac.th
25	มหาวิทยาลัยวลัยลักษณ์	http://www.wu.ac.th
26	มหาวิทยาลัยศรีนครินทรวิโรฒ	http://www.swu.ac.th

ตารางที่ 8 (ต่อ)

ลำดับ	รายชื่อมหาวิทยาลัย	URL
27	มหาวิทยาลัยศรีปทุม	http://www.spu.ac.th
28	มหาวิทยาลัยศิลปากร	http://www.su.ac.th
29	มหาวิทยาลัยสงขลานครินทร์	http://www.psu.ac.th
30	มหาวิทยาลัยสยาม	http://www.siamu.ac.th
31	มหาวิทยาลัยสุโขทัยธรรมาธิราช	http://www.stou.ac.th
32	มหาวิทยาลัยหอการค้าไทย	http://www.utcc.ac.th
33	มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ	http://www.hcu.ac.th

ผลการทดลองแบ่งออกเป็น 3 ส่วนคือ

1. การรวบรวมเว็บเพจที่ประกอบด้วยข้อมูลบุคคล
2. การรวมข้อมูลของบุคคลคนเดียวกันเข้าด้วยกัน
3. การสืบค้นข้อมูล

การรวบรวมเว็บเพจที่ประกอบด้วยข้อมูลบุคคล

ในการทดลองได้มีการนำรายชื่อมหาวิทยาลัยจำนวน 33 มหาวิทยาลัย ดังที่กล่าวมา นำไปจัดเก็บในระบบเสิร์ชเอนจิ้นที่มีชื่อว่า mnoGoSearch ด้วยวิธีการทำอินเด็กซ์ (index) เหตุผลที่เลือกใช้เสิร์ชเอนจิ้นนี้ เนื่องจากมีความสามารถในการตัดคำสำหรับเอกสารภาษาไทย โดยหลังที่เสิร์ชเอนจิ้นได้ทำการรวบรวมข้อมูลดังกล่าวด้วยวิธีการทำอินเด็กซ์เป็นที่เรียบร้อย คำค้นที่เป็นคำนำหน้าชื่อของบุคคลจะถูกส่งเข้าไปในเสิร์ชเอนจิ้นเพื่อนำเว็บเพจที่มีข้อมูลบุคคลออกมา โดยคำนำหน้าชื่อบุคคลที่ใช้ในการทดลองแสดงดังที่แสดงในตารางที่ 9

ตารางที่ 9 คำนำหน้าชื่อบุคคลที่ใช้ในการทดลอง

ลำดับที่	คำนำหน้าชื่อ	ลำดับที่	คำนำหน้าชื่อ	ลำดับที่	คำนำหน้าชื่อ
1	จ.ต.	41	พล.ต.อ.	81	หม่อมหลวง
2	จ.ท.	42	พลตำรวจ	82	เจ้าอธิการพลเอก
3	จ.อ.	43	พลตำรวจตรี	83	พล.อ.

ตารางที่ 9 (ต่อ)

ลำดับที่	คำนำหน้าชื่อ	ลำดับที่	คำนำหน้าชื่อ	ลำดับที่	คำนำหน้าชื่อ
4	จ่าอากาศตรี	44	พลตำรวจเอก	84	พลโท
5	จ่าอากาศเอก	45	พลตำรวจโท	85	พล.ท.
6	จ่าอากาศโท	46	พลฯ	86	พลตรี
7	น.ต.	47	พันตำรวจตรี	87	พล.ต.
8	น.ท.	48	พันตำรวจเอก	88	พันเอก
9	น.อ.	49	พันตำรวจโท	89	พ.อ.
10	นาวาอากาศตรี	50	ร.ต.ต.	90	พันโท
11	นาวาอากาศเอก	51	ร.ต.ท.	91	พ.ท.
12	นาวาอากาศโท	52	ร.ต.อ.	92	พันตรี
13	พ.อ.ต.	53	ร้อยตำรวจตรี	93	พ.ต.
14	พ.อ.ท.	54	ร้อยตำรวจเอก	94	ร้อยเอก
15	พ.อ.อ.	55	ร้อยตำรวจโท	95	ร.อ.
16	พล.อ.ต.	56	ส.ต.ต.	96	ร้อยโท
17	พล.อ.ท.	57	ส.ต.ท.	97	ร.ท.
18	พล.อ.อ.	58	ส.ต.อ.	98	ร้อยตรี
19	พลทหาร	59	สิบตำรวจตรี	99	ร.ต.
20	พลอากาศตรี	60	สิบตำรวจเอก	100	จ่าสิบเอก
21	พลอากาศเอก	61	สิบตำรวจโท	101	จ.ส.อ.
22	พลอากาศโท	62	นาง	102	จ่าสิบโท
23	พลฯ	63	นางสาว	103	จ.ส.ท.
24	พันจ่าอากาศตรี	64	นาย	104	จ่าสิบตรี
25	พันจ่าอากาศเอก	65	บาทหลวง	105	จ.ส.ต.
26	พันจ่าอากาศโท	66	พระ	106	สิบเอก
27	ร.ต.	67	พระครูธรรมธร	107	ส.อ.
28	ร.ท.	68	พระครูปลัด	108	สิบโท
29	ร.อ.	69	พระครูวินัยธร	109	ส.ท.
30	เรืออากาศตรี	70	พระครูสมุห์	110	สิบตรี
31	เรืออากาศเอก	71	พระครูใบฎีกา	111	ส.ต.
32	เรืออากาศโท.	72	พระปลัด	112	พลทหาร
33	จ่าสิบตำรวจ	73	พระมหา	113	ผศ.
34	ด.ต.	74	พระสมุห์	114	ผู้ช่วยศาสตราจารย์
35	นายดาบตำรวจ	75	พระอธิการ	115	รองศาสตราจารย์
36	พ.ต.ต.	76	พระใบฎีกา	116	ศาสตราจารย์
37	พ.ต.ท.	77	ม.ร.ว.	117	อาจารย์
38	พ.ต.อ.	78	ม.ล.	118	รศ.

ตารางที่ 9 (ต่อ)

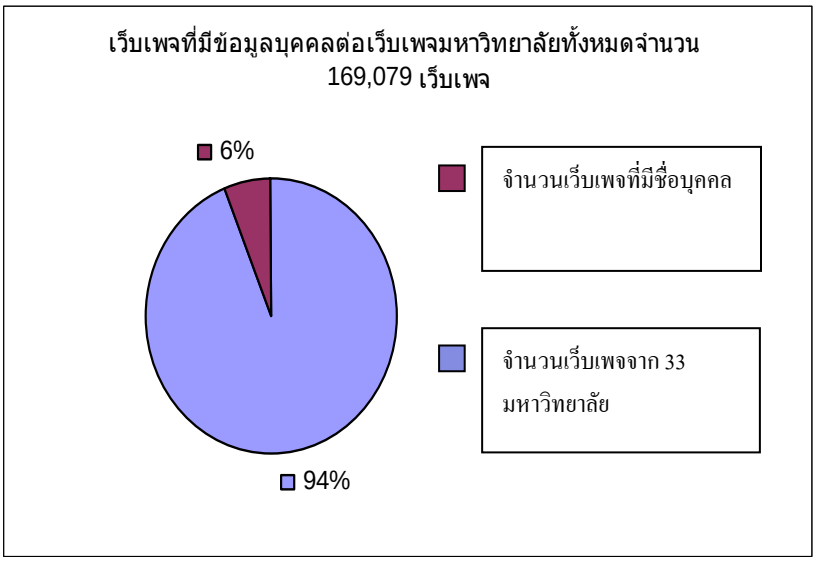
ลำดับที่	คำนำหน้าชื่อ	ลำดับที่	คำนำหน้าชื่อ	ลำดับที่	คำนำหน้าชื่อ
39	พล.ต.ต.	79	สามเณร	119	ดร.
40	พล.ต.ท.	80	หม่อมราชวงศ์	120	รศ.ดร.

แต่ละคำนำหน้าชื่อดังที่แสดงในตารางที่ 9 จะถูกใช้เป็นคำค้นไปยังเสิร์ชเอนจิ้นด้วยโปรแกรมที่ได้พัฒนาขึ้น หลังจากนั้น URL ที่เป็นคำตอบจากเสิร์ชเอนจิ้นส่งกลับค้นมาจะถูกรวบรวมไว้ทั้งหมด ด้วยโปรแกรมที่ได้พัฒนาขึ้นอีกเช่นกัน ซึ่งผลที่ได้แสดงดังตารางที่ 10

ตารางที่ 10 ผลลัพธ์ที่ได้จากการรวบรวมข้อมูลบุคคล

จำนวนมหาวิทยาลัย	จำนวนเว็บเพจทั้งหมด	จำนวนเว็บเพจที่มีข้อมูลบุคคล
33	169,079	11,050

จากตารางที่ 10 นำข้อมูลมาแสดงในรูปแบบของกราฟเพื่อให้เห็นอัตราส่วนระหว่างจำนวนเว็บเพจทั้งหมดต่อจำนวนเว็บเพจที่มีข้อมูลบุคคลดังภาพที่ 29



ภาพที่ 29 อัตราส่วนของจำนวนเว็บเพจของมหาวิทยาลัยทั้งหมดต่อเว็บเพจที่มีข้อมูลบุคคล

1. <u>การรวมข้อมูลของบุคคลคนเดียวกันเข้าด้วยกัน</u>

ในการรวบรวมข้อมูลของบุคคลคนเดียวกันเข้าด้วยกันนั้นได้ทำการทดลองกันข้อมูลรายชื่อคำที่มักจะเขียนผิด ซึ่งมีที่มาจากสารานุกรมฉบับภาษาไทยวิกิพีเดีย ซึ่งลักษณะข้อมูลที่ใช้ทดลองแสดงในภาคผนวก ข

ผลการทดลองของการรวบรวมข้อมูลบุคคลคนเดียวกันเข้าด้วยกันแสดงในตารางที่ 11

<u>ตารางที่ 11</u> ผลการทดลองของการรวมข้อมูลบุคคลคนเดียวกันเข้าด้วยกัน

ค่าขอบเขต	Precision (%)	Recall (%)
1	89.00	71.23
1.5	85.65	88.25
2	77.00	89.40
2.5	64.75	90.10

จากตารางที่ 11 แสดงการวัดประสิทธิภาพของการรวมข้อมูลของบุคคลคนเดียวกันเข้าด้วยกัน การวัดประสิทธิภาพนั้นจะวัดโดยการวัดค่าความถูกต้อง (Precision) และค่าระลึกได้ (Recall) ซึ่งในการทดลองนั้นจะใช้ค่าขอบเขตที่แตกต่างกัน ซึ่งสาเหตุที่ต้องระบุค่าขอบเขตที่แตกต่างกัน เนื่องจากไม่ทราบว่าผลลัพธ์หรือคะแนนที่เกิดจากการคำนวนนี้ค่าใดควรเป็นขอบเขตเพื่อแสดงคู่ของสตริงที่มีค่าเกินจากค่าขอบเขตนี้เป็นคำที่มีความแตกต่างกัน โดยจากผลการทดลองค่าที่ดีที่สุดคือ 1.5

2. <u>การสืบค้นข้อมูล</u>

สำหรับการทดลองการสืบค้นข้อมูลนั้นได้ทดลองกับข้อมูลบุคคลที่สกัดจากเว็บเพจได้จากหน่วยงานมหาวิทยาลัยจำนวน 50 คน โดยในการวัดประสิทธิภาพนั้นจะวัดการสืบค้นจากข้อมูลที่ได้จัดเตรียมไว้ เปรียบเทียบกับการสืบค้นโดยใช้กฎการอนุมานและออนโทโลจีร่วมด้วย ซึ่งออนโทโลจีที่ใช้เป็นออนโทโลจีเกี่ยวกับสาขาความเชี่ยวชาญซึ่งบ่งบอกความสัมพันธ์ระหว่างวิชาในสาขาคอมพิวเตอร์ กฎในการอนุมานที่ได้แสดงรายละเอียดไว้ในหัวข้อวิธีการ จากการทดลองวัดค่าความถูกต้องและค่าระลึกได้ ทำให้ได้ผลลัพธ์ดังตารางที่ 12

ตารางที่ 12 ผลการทดลองการสืบค้นข้อมูล

ประเภทของการสืบค้น	ค่าความถูกต้อง	ค่าความระลึกได้
การสืบค้นบนข้อมูลปกติ	0.79	0.63
การสืบค้นแบบเพิ่มออนโทโลจีและการอนุมาน	0.83	0.85

วิจารณ์

กระบวนการตรวจสอบความคล้ายของชื่อบุคคลนั้น ได้ค่าความถูกต้อง= 0.85 และค่าระลึกได้ = 0.88 โดยความผิดพลาดเกิดจาก 2 สาเหตุได้แก่

ความผิดพลาดของของการระบุว่าตัวอักษรแต่ละตัวทำหน้าที่อะไร เนื่องจากพยางค์บางพยางค์มีความขัดแย้งต่อกฎโดยทั่วไปในการระบุหน้าที่ให้กับแต่ละอักษรที่ได้ออกแบบไว้ ยกตัวอย่างเช่น “ศิลปอาชา” เมื่อแบ่งพยางค์แล้วจะได้ “ศิลป-อา-ชา” โดย “ศิลป” เมื่อเข้าสู่ขั้นตอนในการกำหนดคุณสมบัติให้แต่ละอักขระตามกฎที่ได้ระบุไว้นั้น “ป” จะทำหน้าที่เป็นตัวสะกด ตัวอย่างเช่น เทียบกับชื่อว่า “วิมล” ซึ่ง “ล” อยู่ตำแหน่งเดียวกับ “ป” ในพยางค์ “ศิลป” แต่ความจริงแล้ว “ป” ทำหน้าที่เป็นพยัญชนะต้นเนื่องจากอ่านว่า “ศิน-ละ-ปะ” ไม่ใช่ “ศิ-ลบ” ดังนั้นจึงทำให้การกำหนดคุณสมบัติให้กับอักษร “ป” มีความผิดพลาด เมื่อนำไปเทียบกับค่าน้ำหนักที่กำหนดไว้ก็ทำให้ได้ค่าที่คลาดเคลื่อน เป็นผลทำให้คะแนนที่ได้มาเพื่อใช้ระบุความคล้ายกันของสตริงมีความผิดพลาด

ความผิดพลาดอีกหนึ่งประการคือ รูปแบบของความสับสนในการเขียนมีมากเกินไป ทำให้ไม่สามารถระบุได้ว่าเป็นชื่อของบุคคลเดียวกัน ยกตัวอย่างเช่น “วีรวุฒิ” กับ “วีรวุทธ์” ซึ่งเป็นชื่อที่สามารถเขียนได้ทั้งสองแบบ แต่ความยาวของคำทำให้ค่าคะแนนคือ 3 อีกตัวอย่างหนึ่งคือ “ธงชัย” กับ “ธงไชย” เป็นต้น ทำให้ชื่อเหล่านี้ไม่ผ่านการตรวจสอบค่าคล้าย

ในการทดลองการสืบค้นข้อมูลบุคคลจากสาขาวิชาวิศวกรรมคอมพิวเตอร์โดยอาศัยออนโทโลจีของสาขาความเชี่ยวชาญและกฎการอนุมานแล้วพบว่า ได้ค่าความถูกต้อง0.83 และค่าระลึกได้ 0.85 เมื่อเปรียบเทียบกับการสืบค้นโดยไม่อาศัยออนโทโลจีและกฎการอนุมานแล้ว พบว่ามีค่าความถูกต้องมีค่าเพิ่มขึ้น 0.04 และค่าระลึกได้เพิ่มขึ้น 0.22 สำหรับโดยความผิดพลาดที่เกิดจากการสืบค้นโดยอาศัยกฎการอนุมานและออนโทโลจี เกิดจากออนโทโลจีมีขนาดไม่ครอบคลุมคำค้นที่ผู้ใช้เลือกใช้

สรุป

ในงานวิจัยนี้ มุ่งเน้นการพัฒนาระบบสืบค้นข้อมูลบุคคล โดยได้แบ่งส่วนประกอบของระบบสืบค้นข้อมูลบุคคลออกเป็น 4 ส่วนด้วยกันคือ การรวบรวมข้อมูลบุคคลจากอินเทอร์เน็ต การสกัดข้อมูลบุคคลจากเว็บเพจ การรวมข้อมูลบุคคลเข้าด้วยกัน และการจัดเก็บข้อมูลบุคคล โดยสิ่งที่งานวิจัยนี้ได้พัฒนาเพิ่มขึ้นมานั้นมี 3 ส่วนคือ การรวบรวมข้อมูลบุคคล ได้พัฒนาโดยใช้เสิร์ชเอนจิ้นในการรวบรวมเว็บเพจที่เกี่ยวข้องกับบุคคล ซึ่งทำให้ได้ประหยัดเวลาในการรวมข้อมูลได้มากขึ้นเนื่องจากไม่ต้องเสียเวลาในการกรองเว็บเพจที่ไม่มีข้อมูลบุคคลออกไป ถัดไปคือการรวบรวมข้อมูลบุคคลเข้าด้วยกัน ซึ่งในงานวิจัยนี้ได้รวบรวมโดยใช้การตรวจสอบความคล้ายของชื่อบุคคล โดยได้พัฒนาวิธีการตรวจสอบความคล้ายของสตริงต่อจากวิธีการของ Edit distance ด้วยการเพิ่มค่าน้ำหนักให้กับแต่ละครั้งของการเปลี่ยนแปลง ให้เหมาะสมกับภาษาไทย สุดท้ายได้พัฒนาสืบค้นข้อมูลบุคคล โดยให้มีการจัดเก็บข้อมูลบุคคลในภาษาอาวล์เพื่อรองรับการอนุมานเพื่อทำให้ได้ข้อมูลที่อาจซ่อนอยู่ในข้อมูลจริง และสร้างออนโทโลจีเกี่ยวกับสาขาความเชี่ยวชาญ เพื่อเพิ่มช่องทางในการเข้าถึงข้อมูลของผู้ใช้ให้มากยิ่งขึ้น

การรวบรวมข้อมูลบุคคลนั้นได้มาจากเว็บไซต์ของหน่วยงานมหาวิทยาลัยจำนวน33 มหาวิทยาลัย ซึ่งทำให้ได้เว็บเพจมาจำนวน 169,079 เว็บเพจ และในจำนวนนั้นเป็นเว็บเพจที่มีข้อมูลบุคคลอยู่จำนวนทั้งสิ้น 11,050 เว็บเพจ

การรวมของบุคคลคนเดียวกันเข้าด้วยกัน โดยงานวิจัยนี้ได้รวมข้อมูลเข้าด้วยกันโดยการตรวจสอบความคล้ายกันของชื่อบุคคล วิธีการตรวจสอบความคล้ายที่ได้เลือกใช้นั้นมีพื้นฐานอยู่บนวิธีตรวจสอบความคล้ายกันของสตริงแบบ Edit distance ซึ่งสิ่งที่ได้พัฒนาเพิ่มขึ้นมาเพื่อให้เหมาะสมกับการตรวจสอบความคล้ายของชื่อบุคคลในภาษาไทยคือ การเพิ่มค่าน้ำหนักในการคิดคะแนนเพื่อบ่งบอกความคล้ายกันของสตริง จากผลการทดลองทำให้สามารถสรุปได้ว่า ค่าคะแนนที่สมควรใช้เป็นค่าขอบเขตในการบ่งบอกความคล้ายกันของสตริงคือ 1.5 เนื่องจากค่านี้ทำให้ประสิทธิภาพของการตรวจสอบความคล้ายของชื่อบุคคลมีค่าสูงสุดคือค่าความถูกต้อง 0.85 และ ค่าระลึกได้ 0.88

การสืบค้นข้อมูลบุคคลโดยอาศัยการอนุมานและออนโทโลจีได้ค่าความถูกต้อง = 0.83 และค่าระลึกได้ = 0.85 จึงสรุปได้ว่าการนำวิธีการอนุมานและออนโทโลจีมาประยุกต์ใช้กับการสืบค้นข้อมูลบุคคลช่วยให้ประสิทธิภาพในการสืบค้นดีขึ้น

งานที่ควรปรับปรุงต่อคือ การปรับปรุงวิธีในการแยกส่วนประกอบของพยางค์ ว่าส่วนใดเป็นพยัญชนะต้น สระ ตัวสะกด ซึ่งเป็นส่วนสำคัญที่จะทำให้การตรวจสอบความคล้ายของสตริงมีความถูกต้องมากยิ่งขึ้น อีกส่วนหนึ่งที่สมควรพัฒนาเพิ่มคือการสกัดข้อมูลบุคคลจากเว็บเพจ

เอกสารและสิ่งอ้างอิง

Bing Liu and Chee Wee Chin, **Searching People on the Web According to Their Interrests**, Poster Proc. Of WWW-02, Hawaii, USA, May, 2002

Alexiei Dingli, Fabio Ciravegna, David Guthrie and Yorrick Wilks, **Mining Web Sites Using Unsupervised Adaptive Information Extraction**, In Proceedings 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary, 2003

Aron Culotta, Ron Bekkerman and Andrew McCallum, **Extracting social networks and contact information from email and the web**, Proceedings of CEAS, 2004

Masanori Harada, Shin-ya Sato and Kazuhiro Kazama, **Finding Authoritative People from the Web**, Proceedings of the Joint Conference on Digital Libraries, 2004

Xiaojun Wan, Jianfeng Gao Mu Li, and Binggong Ding, **Person Resolution in Person Search Results: WebHawk**, Proceedings of the 14th ACM international conference 2005

Sylvain Tenier, Amedeo Napoli, Xavier Polanco and Yannick Toussaint, **Knowledge extraction from webpages**, In 5th International Workshop on Knowledge Markup and Semantic Annotation - SemAnnot 2005.

Javier Artiles, Julio Gonzalo and Felisa Verdojo, **A Testbed for People Searching Strategies in the WWW**, Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 2005

Kushmerick, N. Weld, D. S. Doorenbos, R**, Wrapper Induction for Information Extraction**, Lawrence Erlbaum Associate Ltd 1997

R Guha and A Garg ,**Disambiguating People in Search**, Building the Semantic Web, 2003

D Eichmann, **The RBSE Spider -Balancing Effective Search Against Web Load**, Proceedings of the First International World Wide Web, 1994

B Pinkerton, **Finding What People Want: Experiences with the WebCrawler**, Proceedings of the Second International World Wide Web, 1994

OA McBryan, **GENVL and WWW: Tools for Taming the Web,Proc**,. 1st International World Wide Web Conference, 1994

Sergey Brin and Lawrence Page, **The Anatomy of a Large-Scale Hypertextual Web Search Engine** 1998

A Heydon, M Najork, **A scalable, extensible Web crawler,World Wide Web,** Springer,1999

J. Edwards, K. S. McCurley, and J. A. Tomlin, **An adaptive model for optimizing performance of an incremental web crawler**. In Proceedings of the Tenth Conference on World Wide Web, pages 106–113, Hong Kong, May 2001. Elsevier.

Vladislav Shkapenyuk Torsten Suel,**Design and Implementation of a High-Performance Distributed Web Crawler**, Proceeding of the international conference on data, 2002

D Zeinalipour-Yazti, M Dikaiakos, **Design and Implementation of a Distributed Crawler and Filtering Processor,**Lecture notes in computer science, 2002,

P Boldi, B Codenotti, M Santini, S Vigna - **UbiCrawler: a scalable fully distributed Web crawler,** Software- Practice and Experience, 2004

KM Risvik, R Michelsen, - **Search engines and Web dynamics** Computer Networks, 2002

Dussadee Thamvijit, Hutchatai Chanlekha, Chalermpon Sirigayon, Trakul Permpool and Asanee Kawtrakul, "**Person Information Extraction from the web**", The Sixth Symposium on Natural Language Processing 2005 (SNLP 2005), Chiang Rai, Thailand, December 13-15, 2005

Hutchatai Chanlekha and Asanee Kawtrakul, "**Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information**", IJCNLP' 2004 , Hainan Island , China, 2004

VJ Hodge, J Austin, "**An Evaluation of Phonetic Spell Checkers**", 2001

Roger L. Costello,David B. Jacobs, "**A Quick Introduction to OWL Web Ontology Language**" 2001

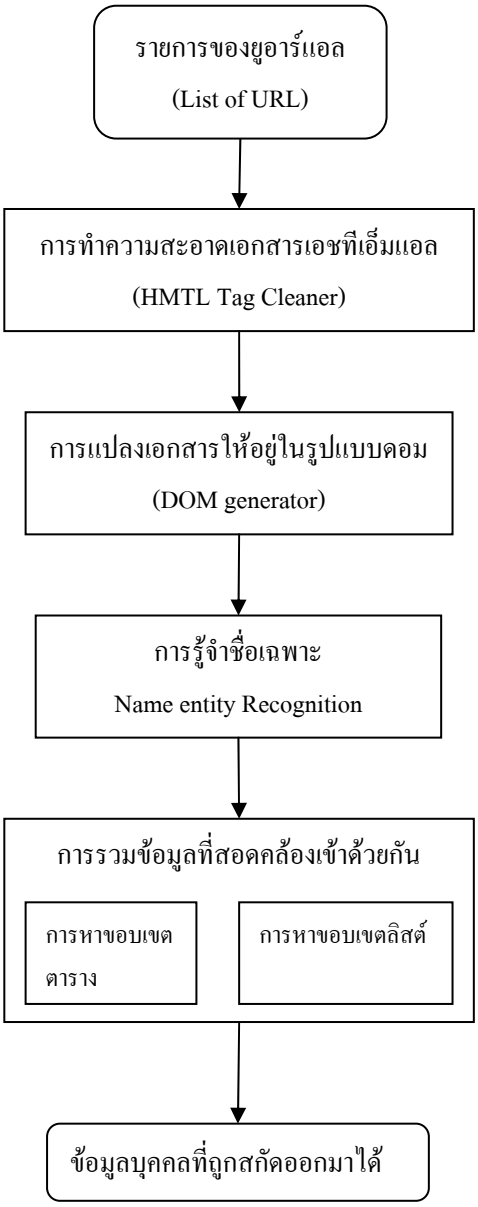
ภาคผนวก

ภาคผนวก ก

ขั้นตอนในการสกัดข้อมูลบุคคล

ขั้นตอนในการสกัดข้อมูล

ในการสกัดข้อมูลบุคคลออกจากเว็บเพจนั้นได้ใช้วิธีการของ(Thamvijit, 2005) ซึ่งขั้นตอนการทำงานของวิธีการดังกล่าวมีขั้นตอนมีทั้งสิ้น 4 ขั้นตอน ได้แก่ การทำความสะอาดเอกสารเอชที่เอ็มแอล, การแปลงเอกสารให้อยู่ในรูปแบบของดอม, การรู้จำชื่อเฉพาะ และการรวมข้อมูลที่สอดคล้องเข้าด้วยกัน ดังแสดงในภาพที่ 30



ภาพที่ 30 แบบจำลองขั้นตอนการทำงานของการสกัดข้อมูลบุคคล

จากภาพที่ 30 ขั้นตอนของการสกัดข้อมูลบุคคลนั้นมีการบวนการย่อยทั้งหมด 4 ขั้นตอน ได้แก่

การทำความสะอาดเอกสารเอชทีเอ็มแอล (HTML tag cleaner)

เว็บเพจนั้นส่วนใหญ่แล้วถูกสร้างขึ้นมาด้วยภาษา HTML แทบทั้งสิ้น และบ่อยครั้งที่โค้ดเอชทีเอ็มแอล ไม่ได้มีความถูกต้อง 100 % ตัวอย่างเช่น ภาพที่ 31

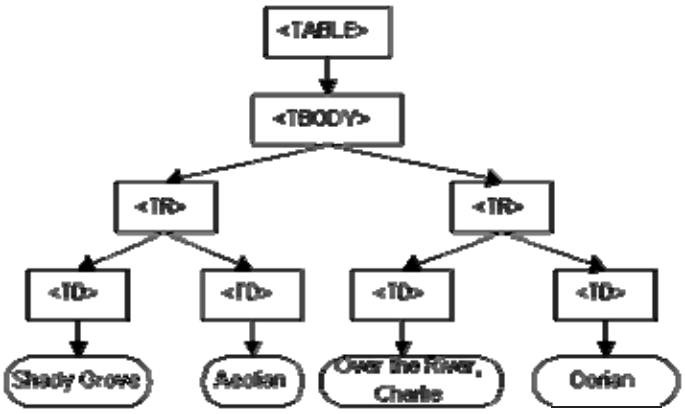
ภาพที่ 31 โค้ดเอชทีเอ็มแอลที่มีความถูกต้องทั้งหมดกับแบบที่ไม่ถูกต้องทั้งหมด

จากภาพที่ 31 ส่วนที่แตกต่างระหว่างโค้ดเอชทีเอ็มแอลที่มีความถูกต้องทั้งหมดกับโค้ดที่ไม่ได้ถูกต้องทั้งหมดนั้นแสดงด้วยการขีดเส้นใต้ โดยลักษณะของโค้ดทั้งสองแบบเมื่อนำเสนอในรูปแบบของเว็บเพจแล้วจะมีความเหมือนกันทุกประการ แต่ในการสกัดข้อมูลบุคคลนั้น ลักษณะของข้อมูลที่จะถูกนำมาประมวลผลนั้นอยู่ในรูปแบบโค้ดเอชทีเอ็มแอล ดังนั้นลักษณะของโค้ดที่มีความผิดพลาดจึงต้องถูกกำจัดออกไป ซึ่งในการกำจัดส่วนที่ผิดนั้นได้ใช้โปรแกรมเจไทดี้ (JTidy: HTML Parser and Pretty Printer in Java)

การแปลงเอกสารให้อยู่ในรูปแบบดอม (DOM: Document Object Model)

เนื่องมาจากข้อมูลที่อยู่ในรูปแบบของเอชทีเอ็มแอลนั้นมีความซับซ้อนและเป็นรูปแบบที่เข้าใจได้ยาก อีกทั้งในการสกัดข้อมูลบุคคลนั้นมุ่งเน้นเพื่อการสกัดข้อมูลบุคคลภายในโครงสร้าง

แบบตารางและลิสต์ ดังนั้นจึงมีความต้องการในการเปลี่ยนแปลงข้อมูลที่อยู่ให้รูปแบบใหม่ โดยมีจุดมุ่งหมายเพื่อความสะดวกในการสกัดข้อมูล ซึ่งวิธีการที่ใช้เพื่อแปลงให้เป็นรูปแบบที่สามารถทำความเข้าใจได้ง่ายขึ้นนั้น เป็นรูปแบบที่สร้างขึ้นโดยองค์กรที่เป็นผู้สร้างภาษาเอชทีเอ็มแอล ซึ่งรูปแบบนั้นมีชื่อว่าดอม (DOM: Document Object Model) โดยลักษณะโดยทั่วไปของดอมนั้นจะแปลงเอกสารเอชทีเอ็มแอลให้อยู่ในรูปแบบของต้นไม้ ดังตัวอย่างในภาพที่ 32

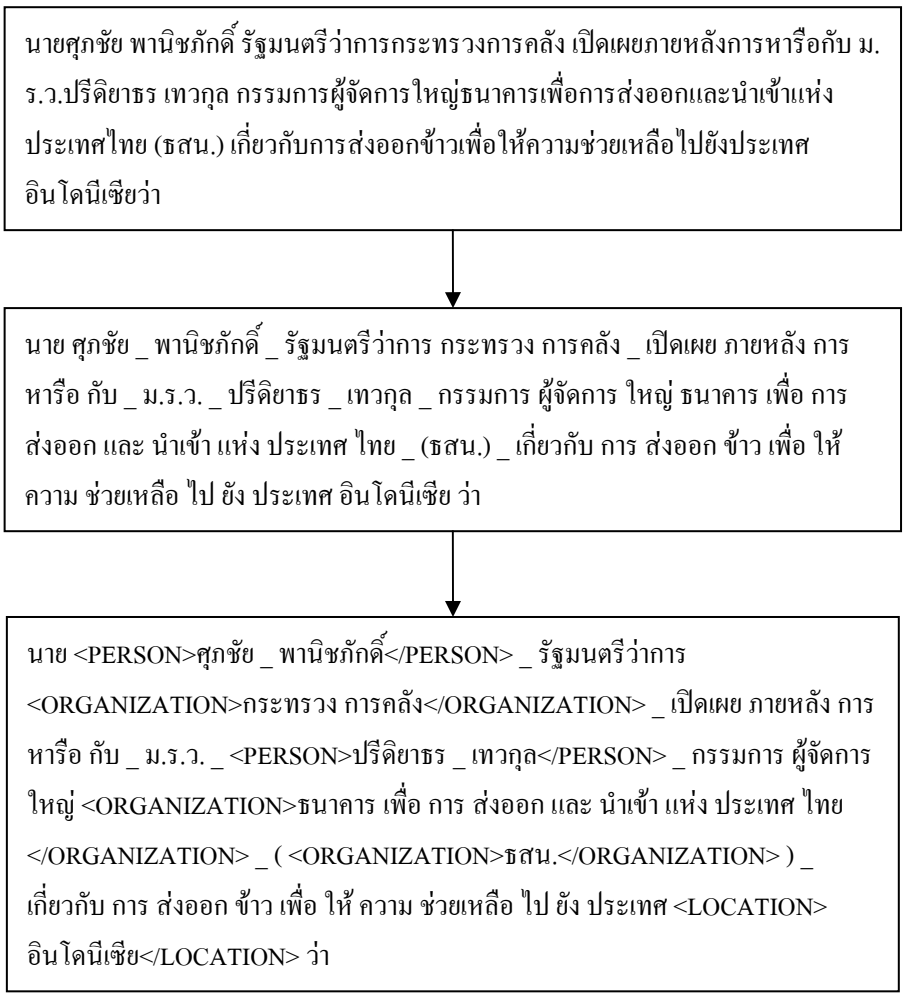


ภาพที่ 32 รูปแบบของเอกสารเอชทีเอ็มแอลที่ถูกจัดเก็บในรูปแบบของดอม

จากภาพที่ 32 แสดงให้เห็นถึงรูปแบบในการจัดเก็บเอกสารเอชทีเอ็มแอลในรูปแบบของดอม ซึ่งจะทำให้สามารถแยกส่วนประกอบต่างๆของเอชทีเอ็มแอล เช่น ตาราง, ลิสต์ เป็นต้น

การรู้จำชื่อเฉพาะ (Name Entity Recognition)

เมื่อเอกสารเอชทีเอ็มแอลถูกแปลงให้อยู่ในรูปแบบดอมแล้ว ก็จะเข้าสู่กระบวนการระบุชื่อเฉพาะ เช่น ชื่อองค์กร, ชื่อบุคคล, ชื่อสถานที่ เป็นต้น โดยจะให้วิธีการระบุชื่อเฉพาะของ (Chanlekha, 2004) ตัวอย่างของการระบุชื่อเฉพาะแสดงในภาพที่ 33



ภาพที่ 33 ลักษณะของการสกัดชื่อเฉพาะ

จากภาพที่ 33 เป็นลักษณะของการสกัดชื่อเฉพาะออกจากเอกสาร ด้วยวิธีการรู้จำชื่อเฉพาะของ (Chanlekha, 2004) โดยเริ่มต้นจากข้อความภาษาไทย ขั้นตอนต่อมาคือการแบ่งขอบเขตของคำ หลังจากนั้นคือการกำกับลงในข้อความว่าส่วนใดคือ ชื่อบุคคล (กำกับด้วย <PERSON>), องค์กร (กำกับด้วย <ORGANIZATION>), สถานที่ (กำกับด้วย <LOCATION>)

<u>การรวมข้อมูลที่เกี่ยวข้องกันเข้าด้วยกัน</u>

เนื่องจากในเว็บเพจหนึ่งๆ อาจมีข้อมูลบุคคลมากเกินกว่า 1 คน ดังนั้นจึงจำเป็นที่จะแยกให้ได้ว่าข้อมูลในแต่ละส่วนเป็นของบุคคลใด ภายหลังจากที่ชื่อเฉพาะต่างๆภายในเว็บเพจถูกระบุเป็นที่เรียบร้อยแล้ว