ระบบสืบค้นข้อมูลแบบบูรณาการเชิงความหมายสำหรับข้อมูลภายในองค์กร

ในปัจจุบันข้อมูลภายในองค์กรมีปริมาณมากขึ้น และถูกจัดเก็บในรูปแบบต่างๆ ทั้งในรูปแบบของฐานข้อมูล เอกสาร อีเมล์ เว็บเพจ ในรูปแบบของระบบอินทราเน็ต (Intranet) จากรายงานของ IDC [1] วิเคราะห์ว่า ผู้ทำงานใช้เวลา 15-35% ของเวลาทำงานไปกับการสืบค้นข้อมูลที่จำเป็นต่อการทำงาน ถึงแม้ว่าข้อมูลสำคัญต่างๆ ขององค์กร มักถูกจัดเก็บลงในระบบอินทราเน็ตแล้ว จากรายงานฉบับดังกล่าวพบว่า 40% ของผู้ทำงานไม่สามารถค้นพบข้อมูลที่ตนเองต้องการใช้งานได้จากระบบอินทราเน็ตขององค์กร

ปัญหาการไม่สามารถค้นพบข้อมูลที่ต้องการได้ในระบบอินทราเน็ต ส่วนหนึ่งเป็นผลมาจากระบบสืบค้นข้อมูลยังไม่สามารถตอบสนองความต้องการของผู้ใช้ได้อย่างเพียงพอ โดยระบบสืบค้นข้อมูลมักเป็นการค้นหาโดยคำสำคัญ (keyword search) กล่าวคือ เมื่อผู้ใช้ป้อน คำค้น (query) ระบบจะค้นคืนเอกสาร หรือค้นหาจากฟิลด์ฐานข้อมูลที่มีคำดังกล่าวปรากฏอยู่ เนื่องจากข้อมูลในระบบอินทราเน็ตนั้นจะเป็นแบบหลากชนิด ทั้งที่อยู่ในรูปแบบของฐานข้อมูล (database), เอกสาร (document), ข้อมูลติดต่อ (contacts), อีเมล์ และเว็บเพจ เป็นต้น การสืบค้นข้อมูลโดยคำสำคัญเพียงช่วยให้ผู้ใช้ค้นพบเอกสารที่มีความเกี่ยวข้อง แต่อาจไม่สามารถค้นพบข้อมูลที่ต้องการได้ เนื่องจากผลลัพธ์ที่ได้ขาดการบูรณาการและเชื่อมโยง ตัวอย่างเช่น ผู้ใช้ต้องการค้นหาว่าบุคคลใดในองค์กรที่มีความเชี่ยวชาญด้าน ? ปัญญาประดิษฐ์? การสืบค้นข้อมูลโดยใช้คำค้นดังกล่าว อาจค้นพบรายการเอกสารที่เกี่ยวข้อง แต่ผู้ใช้อาจไม่ได้รายชื่อของบุคคลากรทั้งหมดที่มีความเชี่ยวชาญสาขาดังกล่าว รวมทั้งไม่มีการบูรณาการเชื่อมโยงกับข้อมูลที่เกี่ยวข้อง เช่น นิยาม (definition) ของคำดังกล่าว ข้อมูลติดต่อของแต่ละบุคคล หรือ เอกสารที่บุคคลเหล่านั้นสร้างขึ้น เป็นต้น

เพื่อให้การสืบค้นข้อมูลจากระบบอินทราเน็ตขององค์กรมีประสิทธิภาพดี สามารถตอบสนองความต้องการข้อมูลของผู้ใช้ได้ดียิ่งขึ้น โครงการนี้จึงนำเสนอการประยุกต์ใช้เทคนิคการสืบค้นข้อมูลที่ใช้การจัดระเบียบข้อมูลเชิงความหมาย (Semantic-based information organization) โดยใช้ประโยชน์จากเทคโนโลยีเชิงความหมาย (Semantic Technology)เพื่อสนับสนุนการบูรณาการและจัดระเบียบข้อมูลเชิงความหมาย ส่งผลให้ให้สืบค้นข้อมูลได้ผลลัพธ์ที่มีการสรุปสาระสำคัญ และมีการเชื่อมโยงระหว่างข้อมูลหลากชนิดมากยิ่งขึ้น

การเพิ่มประสิทธิภาพการสืบค้นข้อมูลภายในองค์กรจำเป็นต้องใช้ประโยชน์จากข้อมูลเชิงความหมาย (Semantic Information) ในหลายรูปแบบมาประกอบกัน ข้อมูลเหล่านี้ได้แก่ เมตาเดตา (Metadata) ข้อมูลเชิงกำกับ (Tag) ซึ่งเป็นคำสำคัญที่กำหนดโดยผู้ใช้ ผสานกับข้อมูลออนโทโลยี (Ontology) เพื่อช่วยให้การเข้าถึงข้อมูลของผู้ใช้ตรงกับความต้องการมากยิ่งขึ้น เครื่องมือการสืบค้นข้อมูลที่ใช้ประโยชน์จากข้อมูลเหล่านี้ ได้แก่ บราวเซอร์เชิงความหมาย (Semantic browser) ที่ช่วยในการท่อง (navigate) และเข้าถึงข้อมูลแบบหลากมิติ (Faceted browsing) เช่น ผู้ใช้สามารถเลือกดูข้อมูลตามกลุ่มของบุคคล (person), หน่วยงาน (organization), โครงการ (project), ผลงานผลิตภัณฑ์ (product), สิ่งตีพิมพ์ (publication), เหตุการณ์ (event) หรือตามความสัมพันธ์ต่างๆได้อย่างชัดเจน เช่น เลือกดูข้อมูล

เฉพาะประเด็นที่เกี่ยวข้องกับ?การประชาสัมพันธ์? ที่ถูกสร้างขึ้นเกี่ยวกับ ?เหตุการณ์? หนึ่งๆ ในช่วง ?สัปดาห์ที่ผ่านมา? โดย ?กลุ่มผู้บริหาร? ขององค์กร? เป็นต้น รูปแบบการค้นหาเชิงความหมาย (Semantic Search) เหล่านี้ ช่วยให้ผู้ใช้สามารถระบุความต้องการข้อมูลของตนเองได้ชัดเจนมากยิ่งขึ้น

รูปแบบการค้นหาข้อมูลบนอินทราเน็ตนั้นจำเป็นต้องมีความครอบคลุมข้อมูลหลากชนิดภายในองค์กร ที่มักประกอบด้วยทั้งข้อมูลชนิดมีโครงสร้าง (structured data) เช่น ฐานข้อมูลต่างๆ และ ข้อมูลชนิดไร้โครงสร้าง (unstructured data) เช่น เอกสาร เว็บเพจ อีเมล์ เป็นต้น โดยความต้องการข้อมูลของผู้ใช้มักเป็นส่วนที่เป็นสาระสำคัญที่อยู่ในข้อมูลเหล่านี้ [2] ได้แก่ ข้อมูลเกี่ยวกับเวลา (When), สถานที่ (Where), เหตุผล (Why), นิยาม (What-is), ผู้เชี่ยวชาญ (Who-knows-about), บุคคล (Who-is), คู่มือปฏิบัติ (How-to), สิ่งที่เกี่ยวข้อง (Tell-me-about) เป็นต้น ดังนั้นการสกัดสาระสำคัญออกจากข้อมูลเหล่านี้ ทั้งในรูปแบบของการสกัดเมตาเดตาอัตโนมัติ (Automatic Metadata Extraction) การรู้จำชื่อเฉพาะต่างๆจากเอกสาร (Name Entity Recognition ? NER) เพื่อนำมาช่วยในการสืบค้นข้อมูลเชิงความหมายจึงมีความสำคัญต่อการเพิ่มประสิทธิภาพการสืบค้นข้อมูลภายในองค์กร

ในโครงการนี้ข้อมูลเมตาเดตาและออนโทโลยีทั้งที่มีการสร้างขึ้นใหม่ และ สร้างโดยอัตโนมัติจะถูกนำเสนอโดยใช้มาตรฐานข้อมูลชนิด RDF และ OWL ตามแนวทางของเว็บเชิงความหมาย (Semantic Web) เพื่อให้เกิดการสร้างและบูรณาการข้อมูลเชิงความหมายในรูปแบบที่เป็นมาตรฐาน รวมทั้งเอื้ออำนวยต่อการสืบค้นข้อมูลอัจฉริยะ (Intelligent search) ผ่านระบบฐานข้อมูล RDF และภาษา SPARQL รวมทั้งสามารถประยุกต์ใช้ระบบอนุมานอัตโนมัติ (Inference Engine) สำหรับเพิ่มประสิทธิภาพการสืบค้นข้อมูลอีกด้วย

<u>วัตถุประสงค์ของโครงการ</u>

- เพื่อพัฒนาระบบสืบค้นค้นข้อมูลบูรณาการเชิงความหมายสำหรับการจัดการข้อมูลและความรู้ภายในองค์กร

- เพื่อพัฒนาเครื่องมือที่เกี่ยวข้องเพื่อเพิ่มประสิทธิภาพการสืบค้นข้อมูลจากแหล่งข้อมูลและความรู้ภายในองค์กร

<u>บทคัดย่อ</u>

การสืบค้นข้อมูลภายในองค์กรหรืออินทราเน็ตมีความสำคัญเป็นอย่างยิ่งในปัจจุบัน เนื่องด้วยปริมาณข้อมูลมหาศาลที่เป็นประโยชน์ต่อบุคคลากรภายในองค์กร ทั้งที่เป็นเอกสาร เว็บเพจ และ ฐานข้อมูลต่างๆ ความต้องการในการสืบค้นข้อมูลในองค์กรมักอยู่ในรูปของ การค้นหานิยามศัพท์ ข้อมูลบุคคลากร ผู้เชี่ยวชาญ ข้อมูลเหตุการณ์ คู่มือการใช้งาน รวมทั้งข้อมูลอื่นๆที่เกี่ยวข้อง เช่น ข้อมูลติดต่อ ที่อยู่โฮมเพจ และ เอกสารอ้างอิง รูปแบบการสืบค้นข้อมูลอินทราเน็ตในปัจจุบัน มักอิงกับการค้นหาโดยใช้คำสำคัญ เพื่อหาคำที่ปรากฏอยู่ในเอกสารและฟิลด์ฐานข้อมูล โดยผู้ใช้ต้องทำการค้นหาข้อมูลแยกตามฐานข้อมูลที่เกี่ยวข้อง โครงการวิจัยนี้เสนอแนวทางใหม่ในการจัดระเบียบและสืบค้นข้อมูลอินทราเน็ตในแบบบูรณาการเชิงความหมาย โดยการสืบค้นข้อมูลจะอิงกับการจัดกลุ่มและความสัมพันธ์ของสิ่งต่างๆในเชิงความหมาย โดยมิได้จำกัดว่าข้อมูลนั้นมาจากฐานข้อมูลเดียว หรือหลายฐานข้อมูล โดยใช้การจัดระเบียบที่อิงกับข้อมูล

เชิงความหมายทั้งที่สกัดจากข้อมูลโดยอัตโนมัติ และสร้างขึ้นเองในรูปแบบของออนโทโลยี ระบบสืบค้นข้อมูลบนอินทราเน็ตนั้นจัดเป็นเครื่องมือสำคัญสำหรับการจัดการความรู้ที่มีอยู่ในองค์กร รวมทั้งเป็นจุดเริ่มต้นที่สามารถนำไปสู่การพัฒนาเว็บเชิงความหมายได้ในอนาคต การพัฒนาระบบสืบค้นข้อมูลแบบบูรณาการเชิงความหมายจำต้องอาศัยการออกแบบข้อมูลเชิงความหมายที่อยู่ในรูปของเมตาเดตาและออนโทโลยีเป็นองค์ประกอบสำคัญในการกำหนดความรู้ของโดเมน เป้าหมายสิ่งส่งมอบของโครงการประกอบไปด้วย 3 ส่วนใหญ่ คือ 1) ระบบสืบค้นข้อมูลแบบบูรณาการเชิงความหมายสำหรับข้อมูลภายในองค์กร 2) ระบบคลังทรัพยากรความรู้สำหรับจัดเก็บเมตาเดตา และออนโทโลยีที่แสดงความสัมพันธ์เชื่อมโยงโดยใช้ภาพ 3) ซอฟแวร์บริการสนับสนุนพื้นฐานต่างๆ เช่น หน่วยรู้จำชื่อเฉพาะ หน่วยรวบรวมจัดเก็บข้อมูลแบบหลากรูปแบบ หน่วยจัดเก็บและค้นคืนข้อมูลตามมาตรฐาน RDF รวมทั้งหน่วยประมวลผลเชิงอนุมาน เป็นต้น

ระยะเวลาดำเนินโครงการ : 5 มกราคม พ.ศ.2552 ถึง 4 มกราคม พ.ศ.2554

คณะผู้วิจัย

หัวหน้าโครงการ : นายมารุต บูรณรัช

ผู้ร่วมวิจัย : นายสภา จรรยาชัชวาล, นางสาวพรพิมล ผลินกูล, นายเทพชัย ทรัพย์นิธิ, นายสุพล ไกลถิ่น