โครงการวิจัยพัฒนาทรัพยากรสำหรับการจัดการข้อมูลขนาดใหญ่บนเครือข่ายอินเทอร์เน็ต การสืบค้นข้อมูลและการขุดหาข้อมูล

(Research and Development of Resources for Processing Very Large-Scaled Information on the InternetInformation Retrieval and Data Mining)

ชื่อผู้วิจัย/หน่วยงาน

ร.ศ.ดร.ธนารักษ์ ธีระมั่นคง — ห้องปฏิบัติการการจัดการข้อมูลสารสนเทศและความรู้ สถาบันเทคโนโลยีนานาชาติสิรินธร มหาวิทยาลัยธรรมศาสตร์

ผ.ศ.ดร.บุญเสริม กิจศิริกุล — ห้องปฏิบัติการอัจฉริยะจักรกลและการค้นพบความรู้ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

บทคัดย่อ

โครงการนี้มีจุดประสงค์ในการพัฒนาฐานข้อมูลและระบบข้อมูลที่ช่วยในการจัดการและนำข้อมูลขนาดใหญ่ในอินเทอร์มาใช้ประโยชน์ โดยใช้เทคโนโลยีภาษามนุษย์ ปัจจุบันข้อมูลส่วนใหญ่ที่อยู่บนอินเทอร์เน็ตนั้น ยังไม่มีการจัดโครงสร้างที่ดีในด้านความหมาย เนื่องจากมีข้อจำกัดทางด้านภาษาที่ใช้ในการแสดงผลบนเว็บ โดยภาษา HTML นั้นไม่มีส่วนของความหมายรวมอยู่ด้วย แม้ว่าปัจจุบันจะมีแนวโน้มในการใช้ภาษาที่มีความหมาย เช่น XML ในการสร้างข้อมูลเว็บ แต่ข้อมูลที่มีประโยชน์บนเว็บส่วนใหญ่ยังอยู่ในรูปของเอกสารข้อความธรรมดา หรือเอกสาร HTML ที่ไม่มีโครงสร้างด้านความหมายในประเทศไทย ยังมีงานวิจัยทางด้านนี้น้อยมาก แม้ว่าเราจะมีข้อมูลอยู่ในรูปของภาษาไทยอยู่ในอินเทอร์เน็ตมากมายก็ตาม ในโครงการนี้ การวิจัยและพัฒนาเทคนิคต่างๆ ในการสกัดสารสนเทศและความรู้จากข้อมูลที่ไม่มีโครงสร้างบนอินเทอร์เน็ต ภารกิจหลักในการวิจัยนี้มี 5 หัวข้อด้วยกันคือ 1) การสร้างทรัพยากรสารสนเทศจากอินเทอร์เน็ต ซึ่งได้แก่ ฐานข้อมูลด้านหมวดหมู่ของเอกสารบนอินเทอร์เน็ตและคำศัพท์ต่างๆ ที่ใช้ในแต่ละหมวดหมู่ 2) การเตรียมเครื่องมือพื้นฐานที่ใช้ในการประมวลผล จัดเก็บ และสกัดสารสนเทศจากข้อมูลเอกสารที่ได้มา เช่น เครื่องมือในการใส่หมวดหมู่ให้กับเอกสารเว็บ เครื่องมือในการตัดคำ เครื่องมือในการสกัดข้อมูลลิงก์ เป็นต้น 3) การพัฒนาระบบค้นคืนสารสนเทศและระบบทำเหมืองเอกสาร ได้แก่ ระบบค้นคืนเอกสารด้วยคำสำคัญ ระบบจัดหมวดหมู่เอกสารอัตโนมัติ ระบบหาความสัมพันธ์เอกสารอัตโนมัติ 4) การเผยแพร่ความรู้ที่ได้จากงานวิจัยในรูปแบบของบทความในวารสารวิชาการและงานประชุมวิชาการระดับนานาชาติ และ 5) การพัฒนาบุคลากรที่มีความเชี่ยวชาญด้านเทคโนโลยีภาษามนุษย์

บทนำ

ปัจจุบันอินเทอร์เน็ตเป็นสื่อที่ใช้กันแพร่หลายทั่วไป ทั้งที่เป็นการสื่อสารทางเดียว เช่น การเผยแพร่ข้อมูลต่างๆ และที่เป็นการสื่อสารสองทาง เช่น การทำธุรกิจบนเครือข่ายอินเทอร์เน็ต เป็นต้น จากความแพร่หลายนี้ จึงทำให้ข้อมูลบนเครือข่ายมีปริมาณมาก ข้อมูลมากมายเหล่านี้สามารถนำมาใช้ประโยชน์ได้หลายด้าน อย่างไรก็ตามการขยายตัวอย่างรวดเร็วของข้อมูลทำให้เราประสบกับปัญหาของการมีข้อมูลที่หลากหลายและซับซ้อนมาก จนทำให้การนำข้อมูลเหล่านี้มาใช้ประโยชน์นั้นเป็นไปได้ยาก การวิจัยพัฒนาระบบและฐานข้อมูลสำหรับจัดการกับข้อมูลขนาดใหญ่บนเครือข่ายจึงเป็นสิ่งที่จำเป็นมาก เนื่องจากข้อมูลต่างๆ บนเครือข่ายอินเทอร์เน็ตส่วนใหญ่อยู่ในรูปแบบของตัวอักษร ซึ่งไม่มีโครงสร้างที่ชัดเจน แต่มีความหมายและสามารถนำมาใช้ประโยชน์ได้มากเทคโนโลยีที่มีความสำคัญมากในงานวิจัยนี้ก็คือ เทคโนโลยีภาษามนุษย์ เทคโนโลยีการประมวลผลภาษาธรรมชาติ และเทคโนโลยีปัญญาประดิษฐ์ ปัจจุบันในประเทศไทย งานวิจัยและพัฒนาเพื่อประยุกต์เทคโนโลยีด้านเหล่านี้มาใช้ในการจัดการข้อมูลขนาดใหญ่บนอินเทอร์เน็ตยังมีน้อยมาก ทั้งที่ข้อมูลภาษาไทยที่อยู่บนเครือข่ายก็มีขนาดใหญ่ขึ้นมากและมีอัตราการขยายตัวสูง ดังนั้นการศึกษาวิธีการจัดการข้อมูลภาษาไทยขนาดใหญ่ และ/หรือ ข้อมูลภาษาไทยผสมผสานกับภาษาอื่น เช่น ภาษาอังกฤษซึ่งมีอยู่มากที่สุดในเครือข่ายอินเทอร์เน็ต จึงเป็นสิ่งที่สำคัญมาก ดังนั้นเพื่อให้บรรลุจุดมุ่งหมายในการวิจัยและพัฒนาระบบที่สามารถจัดการกับข้อมูลขนาดใหญ่บนเครือข่ายอินเทอร์เน็ตทั้งที่เป็นภาษาไทยและภาษาอื่นๆ ได้ โครงงานนี้จึงเน้นการวิจัยพัฒนาทรัพยากรสำหรับการจัดการข้อมูลขนาดใหญ่บนเครือข่ายอินเทอร์เน็ต โดยเฉพาะระบบการสืบค้นข้อมูลและระบบการขุดหาข้อมูล (Information Retrieval/Data Mining)

วัตถุประสงค์

1. เพื่อพัฒนาฐานข้อมูลที่เกี่ยวกับข้อมูลในเครือข่ายอินเทอร์เน็ต โดยฐานข้อมูลนี้จะมีประโยชน์ต่อการวิจัย และพัฒนาระบบต่างๆ ที่เกี่ยวข้องกับข้อมูลอักษรได้ในอนาคต
2. เพื่อนำเทคโนโลยีปัญญาประดิษฐ์มาใช้ในการพัฒนาความสามารถของระบบให้มีความชาญฉลาด
3. เพื่อพัฒนาซอฟต์แวร์พื้นฐานที่สามารถประยุกต์ใช้งานในระบบต่างๆ
4. เพื่อสร้างระบบค้นคืนข้อมูลภาษาไทยพร้อมกับภาษาอื่น
5. เพื่อสร้างระบบสร้างระบบขุดหาความรู้หรือทำเหมืองข้อมูลเอกสาร
6. เพื่อพัฒนาบุคลากรทางด้านสร้างการจัดการข้อมูลขนาดใหญ่บนเครือข่ายอินเทอร์เน็ตและสาขาอื่นๆ ที่เกี่ยวข้อง เช่น เทคโนโลยีภาษามนุษย์ การประมวลภาษาธรรมชาติ ปัญญาประดิษฐ์ การทำเหมืองข้อมูล

ระเบียบวิธีวิจัย

ภาพรวมของฐานข้อมูลและระบบต่างๆ ที่พัฒนาภายใต้โครงการวิจัยพัฒนานี้ แสดงอยู่ในรูปที่ 1 ในขั้นตอนแรกข้อมูลบทความ (เว็บเพจ) ในเครือข่ายอินเทอร์เน็ต จะถูกดึงเข้าสู่ระบบ โดยกระบวนการดึงและเก็บข้อมูลบทความเว็บเพจ (Web Page Retrieval and Storage) ซึ่งจะสามารถดึงบทความ (เว็บเพจ) จากเว็บไซต์ต่างๆ มาสู่ระบบและจัดเก็บไว้ในฐานข้อมูลเว็บเพจ (Web Page Database) ในขบวนการนี้ ระบบจะตั้งต้นจากเว็บไซต์จำนวนหนึ่งที่กำหนดไว้แต่แรก แล้วจะดึงข้อมูลบทความเว็บเพจอื่นๆ โดยตามลิงก์ที่เชื่อมโยงจากเว็บเพจเริ่มต้นเหล่านี้ไปยังเว็บเพจอื่นๆ ขบวนการนี้จะเกิดซ้ำในรูปแบบเดียวกันไปเรื่อยๆ จากฐานข้อมูลเว็บเพจที่ดึงมาได้ จะผ่านขบวนการต่างๆ เพื่อสร้างข้อมูลเพิ่มขบวนการประมวลผลประกอบด้วย 2 ทางหลักๆ คือ

1. การประมวลผลโดยจัดการเอง (Manual Processing) วิธีนี้ผู้พัฒนาระบบจัดการสารสนเทศ จะต้องทำการพัฒนาระบบเอง โดยเว็บเพจที่นำเข้ามานั้น จะถูกนำเข้ากระบวนการกำกับเว็บ (Category Tagger) ซึ่งจะจัดกลุ่มประเภทของเว็บเพจเข้าด้วยกัน พร้อมทั้งมีการสร้างฐานข้อมูลประเภทของเว็บเพจ (Categories Database) เพื่อนำไปใช้งานต่อไป
2. การประมวลผลโดยการจัดการอัตโนมัติ (Automatic Processing) จะสร้างสารสนเทศให้แก่เว็บเพจโดยอัตโนมัติ โดยข้อมูลดิบของเว็บเพจจะถูกนำเข้าสู่กระบวนการหาภาษาของเว็บ (Language Identification) เพื่อทำการระบุภาษาของเว็บเพจนั้นๆ ซึ่งจะมีผลต่อการเลือกวิธีการในการแบ่งพยางค์ และแบ่งคำของข้อมูลในขั้นตอนการตัดพยางค์และตัดคำ (Syllable Segmentation and Word Segmentation) ตามลำดับ โดยกระบวนการนี้อาจมีการใช้ข้อมูลจากดิกชันนารี ช่วยในการตัดคำเพื่อเพิ่มความถูกต้อง และสามารถที่จะสร้างความรู้ด้านการตัดคำ (Segmentation Knowledge) ที่จะช่วยปรับปรุงความสามารถในการตัดคำครั้งต่อไปให้เกิดความถูกต้องมากขึ้น จากนั้นข้อมูลเหล่านี้จะถูกนำเข้าสู่กระบวนการวิเคราะห์เว็บเพจเพื่อวิเคราะห์และจัดกลุ่มความสัมพันธ์ของเว็บเพจ และนำไปจัดประเภทของเว็บเพจ ตามฐานข้อมูลประเภทที่มีอยู่ในกระบวนการจัดหมวดหมู่เอกสาร โดยอัตโนมัติ (Text Classification)

ข้อมูลสารสนเทศของเว็บเพจที่ได้จากวิธีการประมวลผลทั้ง 2 วิธี จะถูกนำไปจัดเก็บโดยสร้างดัชนี เพื่อให้สามารถค้นหาสารสนเทศ ได้อย่างรวดเร็ว นอกจากนี้ยังมีกระบวนการในการเรียนรู้เพื่อปรับปรุงสารสนเทศนั้นให้มีความถูกต้องมากที่สุด สำหรับในส่วนของการใช้งานนั้นได้แบ่งออกเป็น 2 ประเภท คือ (1) งานด้านการค้นคืนสารสนเทศ และ (2) งานทำเหมืองข้อมูลบนเอกสาร โดยงานแรกเป็นการสร้างโปรแกรมค้นหา (Search Engine) เพื่อค้นหาเว็บเพจต่างๆ ตามคำสำคัญที่ใส่เข้าไป ส่วนงานหลัง เป็นการหาวิธีการในการจัดแบ่งหมวดหมู่ข้อมูล หรือระบุภาษาของเว็บ นอกจากนี้ยังรวมไปถึงงานด้านการทำเหมืองข้อมูลเพื่อหาความสัมพันธ์ระหว่างเอกสาร โดยวิธีการขุดหากฎสัมพันธ์ (Association Rule Mining) เพื่อสร้างเป็นสารสนเทศและฐานความรู้ต่อไป

ผลการวิจัย

1. ฐานข้อมูลบนอินเทอร์เน็ตประเภทข้อมูลหมวดหมู่
   - การกำกับหมวดหมู่จำนวน 63,579 เว็บเพจจากที่ดึงมาได้ 178,498 เว็บเพจ
   - การกำกับนามเฉพาะจำนวน 8,483 คำจาก 791 เว็บเพจ
2. ซอฟต์แวร์พื้นฐานที่ใช้ในการประมวลผลข้อมูลขนาดใหญ่บนอินเทอร์เน็ต 8 ชิ้นงาน
   - โปรแกรมช่วยในการจัดหมวดหมู่ข้อมูลบทความเว็บเพจ (Web Page Category Tagger)
   - โปรแกรมแยกแยะอัตโนมัติบทความภาษาไทยกับบทความภาษาอื่น (Thai/Non Thai Web Page Classifier)
   - โปรแกรมหาความสัมพันธ์ระหว่างคำ (Rule Association Miner)
   - โปรแกรมหาคำในบทความแบบทั้งบทความ (Full Text Search) โปรแกรมตัดกลุ่มตัวอักษรแบบ TCC (Thai Character Cluster Segmentation)
   - โปรแกรมแยกข้อมูลลิงก์ (Web Link Information Extraction)
   - โปรแกรมแยกภาษาของเอกสาร (Language Identification)
3. ระบบสืบค้นข้อมูลบนเครือข่ายอินเทอร์เน็ต
   - ระบบสืบค้นโดยนำโครงมาจากฟรีซอฟต์แวร์ชื่อ mnoGoSearch มาปรับปรุงให้สามารถรองรับภาษาไทย โดยเชื่อมต่อกับฐานข้อมูล Postgres บนระบบปฏิบัติการ Linux และพัฒนา GUI ให้ใช้งานได้
4. ระบบขุดค้นข้อมูลบนเครือข่ายอินเทอร์เน็ต
   - โปรแกรมจัดหมวดหมู่เอกสารอัตโนมัติด้วยวิธีการเซ็นทรอย
   - โปรแกรมจัดหมวดหมู่เอกสารอัตโนมัติด้วยวิธีการสอนใบวับบวนซ้ำ
   - ระบบหาความสัมพันธ์ระหว่างเอกสารโดยวิธีการทางการทำเหมืองข้อมูลแบบการขุดค้นกฎสัมพันธ์
5. บทความทางวิชาการ/สิ่งตีพิมพ์ในหรือต่างประเทศ จำนวน 26 ชิ้นงาน
   - บทความในวารสารระดับนานาชาติ (International Journal) จำนวน 8 ชิ้นงาน
   - บทความในหนังสือรวบรวมบทความ (Book Chapter/Lecture Note) จำนวน 3 ชิ้นงาน
   - บทความในงานประชุมวิชาการระดับนานาชาติ (International Conference) จำนวน 14 ชิ้นงาน
   - บทความในงานประชุมวิชาการระดับในประเทศ (National Conference) จำนวน 1 ชิ้นงาน
6. จำนวนบุคลากรที่พัฒนา
   - จำนวน 6 คนระดับปริญญาโท เอก (จบโท 4 คน จบเอก 2 คน)

โครงการนี้ได้รับทุนอุดหนุนการวิจัยและพัฒนา จาก ฝ่ายเครือข่ายการวิจัยและพัฒนา (RDD) ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ
112 อุทยานวิทยาศาสตร์ประเทศไทย ถนนพหลโยธิน
ตำบลคลองหนึ่ง อำเภอคลองหลวง จังหวัดปทุมธานี 12120
โทรศัพท์ 02-564-6900 ต่อ 2501-10 โทรสาร 02-564-6901..2

http://www.nectec.or.th/