การตัดคำภาษาไทยสำหรับการค้นคืนข้อมูลด้านเทคโนโลยีสารสนเทศ

นายโกญจนพงษ์ ทองเพชร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

มหาวิทยาลัยเทคโนโลยีสุรนารี

ปีการศึกษา 2555

A THAI WORD SEGMENTATION FOR THE IT DATA RETRIEVAL

Konjanapong Thongphet

A Thesis Submitted in Partial Fulfillment of the Requirements for the

Degree of Master of Engineering in Computer Engineering

Suranaree University of Technology

Academic Year 2012

การตัดคำภาษาไทยสำหรับการค้นคืนข้อมูลด้านเทคโนโลยีสารสนเทศ

มหาวิทยาลัยเทคโนโลยีสุรนารี อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

คณะกรรมการสอบวิทยานิพนธ์

_____

(รศ. ดร. กิตติศักดิ์ เกิดประสพ)
ประธานกรรมการ

_____

(ผศ. ดร. คะชา ชาญศิลป์)
กรรมการ (อาจารย์ที่ปรึกษาวิทยานิพนธ์)

_____

(ผศ. ดร. ปรเมศวร์ ห่อแก้ว)
กรรมการ

_____

(ศ. ดร. ชูกิจ ลิมปิจำนง)
รองคณบดีฝ่ายวิชาการ

_____

(รศ. รอ. ดร. กนต์ธร ชำนิประศาสน์)
คณบดีสำนักวิชาวิศวกรรมศาสตร์

โกญจนพงษ์ ทองเพชร : การตัดคำภาษาไทยสำหรับการค้นคืนข้อมูลด้านเทคโนโลยีสารสนเทศ (A THAI WORD SEGMENTATION FOR THE IT DATA RETRIEVAL) อาจารย์ที่ปรึกษา : ผศ. ดร.คะชา ชาญศิลป์, 96 หน้า.

ข่าวสารข้อมูลสารสนเทศต่างๆ ในรูปแบบอิเล็กทรอนิกส์มีเพิ่มขึ้นมากมายในปัจจุบัน จึงได้มีการพัฒนาระบบค้นคืนข้อมูลขึ้นมาใช้เพื่อให้ได้ข้อมูลที่ถูกต้องที่สุด แต่ในการค้นคืนข้อมูลในภาษาไทยนั้นยังไม่มีการพัฒนาให้สามารถค้นคืนผลลัพธ์ได้ดีเท่าที่ควร เนื่องจากรูปแบบการเขียนภาษาไทยนั้นไม่มีจุดสิ้นสุดคำที่แน่นอน ทำให้การค้นหาแบบการอ้างอิงคำหลัก (Keyword-Base) ทำได้ไม่ดีนัก จากปัญหานี้ผู้วิจัยจึงได้เสนอแนวคิดในการนำเทคโนโลยีเชิงความหมาย (Semantic Technology) มาใช้งานในการพัฒนาระบบค้นคืนข้อมูลภาษาไทยและใช้ออนโทโลจี (Ontology) เข้ามาอธิบายความสัมพันธ์ของคำต่างๆในภาษาไทย ซึ่งมีการทำงานในลักษณะของเว็บแอพพลิเคชัน

สาขาวิชาวิศวกรรมคอมพิวเตอร์ ลายมือชื่อนักศึกษา_____

ปีการศึกษา 2555 ลายมือชื่ออาจารย์ที่ปรึกษา_____

KONJANAPONG THONGPHET : A THAI WORD SEGMENTATION FOR THE IT DATA RETRIEVAL. THESIS ADVISOR : ASST. PROF. KACHA CHANSILP, Ph.D., 96 PP.

ONTOLOGY/ SEMANTIC TECHNOLOGY/ DATA RETRIEVAL

Nowadays, as the amount of electronic information technology are getting more and more. Therefore, the information retrieval system was used to obtain the most accurate information. However, the data retrieval in Thai is not developed to be able to retrieve the results as good as it should. The Keyword-base is not effective enough. Because, in Thai that have no explicit boundary delimiter. Therefore, researcher has designed Thai Data Retrieval by using Semantic Technology, and use Ontology Technology to describe relationship between Thai words that work on web application.

School of Computer Engineering Student's Signature_____

Academic Year 2012 Advisor's Signature_____

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงด้วยดี เนื่องจากได้รับความช่วยเหลืออย่างดียิ่ง ทั้งด้านวิชาการและด้านการดำเนินงานวิจัยจากบุคคลและกลุ่มบุคคลต่างๆ ได้แก่

ผู้ช่วยศาสตราจารย์ ดร.คะชา ชาญศิลป์ อาจารย์ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ให้โอกาสและคำแนะนำช่วยแก้ปัญหามาโดยตลอด รวมทั้งช่วยตรวจทานแก้ไขวิทยานิพนธ์เล่มนี้จนสมบูรณ์

รองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ หัวหน้าสาขาวิชาวิศวกรรมคอมพิวเตอร์ รองศาสตราจารย์ ดร.นิตยา เกิดประสพ ผู้ช่วยศาสตราจารย์ สมพันธุ์ ชาญศิลป์ ผู้ช่วยศาสตราจารย์ ดร.พิชโยทัย มหัทธนาพิวัฒน์ ผู้ช่วยศาสตราจารย์ ดร.ปรเมศวร์ ห่อแก้ว และอาจารย์ ดร.ชาญวิทย์ แก้วกสิ ที่ให้คำปรึกษามาโดยตลอด

ขอขอบคุณ นายสมคะเน บาลลา นักศึกษาปริญญาโท สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี ที่ให้คำปรึกษาด้านการพัฒนาระบบ นายสมชาย สุขอินทร์ นักศึกษาปริญญาเอก สาขาวิชาคณิตศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีสุรนารี ที่ให้คำปรึกษาในการค้นคว้าข้อมูลอ้างอิงที่เกี่ยวข้องกับงานวิจัยนี้

สุดท้ายนี้ผู้วิจัยขอขอบคุณ บิดา มารดา พี่สาวของผู้วิจัย ตลอดจนครูอาจารย์ที่เคารพทุกท่านที่ได้ประสาทความรู้และประสบการณ์ที่ดีมาโดยตลอด

โกญจนพงษ์ ทองเพชร

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหา

ในปัจจุบันข้อมูลสารสนเทศต่างๆ เป็นส่วนสำคัญที่ทำให้บุคลากรหรือหน่วยงานต่างๆ ทำงานได้อย่างมีประสิทธิภาพ โดยสารสนเทศต่างๆ นั้นสามารถนำมาใช้อ้างอิงหรือนำมาเปรียบเทียบกับสารสนเทศจากหลายๆ แหล่ง เพื่อให้สามารถเพิ่มประสิทธิภาพในการทำงานได้อย่างหลากหลายรูปแบบ แต่เนื่องจากมีการเพิ่มขึ้นของจำนวนข้อมูลสารสนเทศที่รวดเร็ว ทำให้ปัญหาที่เกิดตามมาคือ การค้นคืนสารสนเทศเพื่อคัดกรองข้อมูลสารสนเทศที่ต้องการนำมาใช้ขาดความแม่นยำในการแสดงผลลัพธ์หรืออาจได้ผลลัพธ์ที่มีจำนวนมากเกินไป ไม่ครอบคลุมตามความต้องการของผู้ค้นหา เนื่องจากการค้นคืนสารสนเทศส่วนใหญ่ในปัจจุบันมีวิธีการค้นคืนโดยใช้หลักการค้นคืนสารสนเทศจากเอกสารที่สอดคล้องหรือเหมือนคำหลัก (Keyword Matching) แต่ในการค้นคืนสารสนเทศในชีวิตประจำวันนั้น ผู้ใช้งานอาจมีความต้องการค้นคืนสารสนเทศซึ่งเป็นเอกสารที่เฉพาะเจาะจงในส่วนต่างๆ เช่น ในส่วนของเนื้อหาหรือชื่อเจ้าของผลงาน โดยเฉพาะอย่างยิ่งในการค้นคืนสารสนเทศด้วยภาษาไทยยิ่งมีความแม่นยำของผลลัพธ์ในการค้นหาที่ต่ำ เนื่องด้วยเพราะรูปแบบการเขียนในภาษาไทยมีความซับซ้อนไม่มีการกำหนดจุดหยุดของคำที่แน่นอน อีกทั้งในคำที่เขียนตามคำออกเสียงภาษาอังกฤษ มักมีความเข้าใจที่ไม่ตรงกันในรูปแบบของการเขียนภาษาไทยและคำบางคำยังสามารถมีได้หลายความหมาย ทำให้ผลลัพธ์ในการค้นคืนสารสนเทศขาดความถูกต้องแม่นยำ และทำให้ผู้ใช้งานเสียเวลาในการคัดกรองข้อมูลเพื่อนำไปใช้งานจากผลลัพธ์จำนวนมากที่ค้นคืนได้ ผู้วิจัยจึงเกิดแนวคิดว่าถ้าระบบช่วยค้นคืนสารสนเทศสามารถเข้าใจความหมายของคำหลักในการค้นคืนสารสนเทศภาษาไทยได้ตามที่ผู้ค้นคืนต้องการ จะทำให้การค้นคืนสารสนเทศเป็นไปได้อย่างสะดวกและรวดเร็วยิ่งขึ้น เนื่องจากระบบค้นคืนและผู้ใช้งานเข้าใจในคำหลักของการค้นคืนตรงกัน ผลลัพธ์ที่ได้จะเป็นข้อมูลที่ครอบคลุมและเจาะจงมากขึ้นทำให้ผู้ใช้งานใช้เวลาในการคัดกรองข้อมูลจากผลลัพธ์ที่ได้น้อยลง สามารถทำให้ประสิทธิภาพของระบบค้นคืนสารสนเทศเพื่อการใช้งานเพิ่มขึ้น

ผู้วิจัยได้ทำการศึกษาถึงวิธีแก้ปัญหาของระบบให้บริการค้นคืนสารสนเทศ เพื่อมุ่งหวังให้แสดงผลลัพธ์ในการค้นคืนสารสนเทศได้ตรงตามที่ผู้ใช้งานต้องการมากที่สุด โดยจะได้นำหลักการ

เว็บเชิงความหมายเข้ามาช่วยพัฒนาระบบค้นคืนสารสนเทศ โดยจะมุ่งเน้นไปที่การค้นคืนข้อมูลด้านเทคโนโลยีสารสนเทศที่เป็นภาษาไทย ซึ่งการนำหลักการเชิงความหมาย (Semantic) มาใช้งานในการจัดความสัมพันธ์ระหว่างความหมายของคำให้ถูกต้องมากขึ้น โดยการวิเคราะห์คำแต่ละคำที่มีความหมายเหมือนกันหรือคล้ายคลึงกัน นำมาจัดเป็นกลุ่มความสัมพันธ์ของคำและจะใช้รูปแบบการตัดคำในหลายๆ รูปแบบมาทดสอบในการตัดคำหลักที่ผู้ใช้ระบุมา เพื่อคัดเลือกการตัดคำที่ถูกต้องตามหลักการมากที่สุดเพื่อนำมาใช้ในการพัฒนาระบบค้นคืนสารสนเทศ เนื่องจากคำหลักที่ถูกต้องตามที่ผู้ใช้งานระบุ จะทำให้การค้นคืนสารสนเทศได้ผลลัพธ์ที่แม่นยำขึ้นและใช้ในการตัดคำหรือแบ่งประโยคของเนื้อหาในเอกสารที่เป็นข้อมูลของการค้นคืน เพื่อนำมาใช้สร้างเป็นดัชนีของการค้นคืน ส่วนการค้นคืนสารสนเทศที่เป็นภาษาอังกฤษจะอ้างอิงจากความสัมพันธ์ของข้อมูลจากฐานข้อมูลเวิร์ดเน็ต (WordNet) ซึ่งเป็นฐานข้อมูลเชิงความหมายของกลุ่มคำภาษาอังกฤษที่ได้รับการยอมรับมากที่สุด

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาโครงสร้างการทำงานของระบบค้นคืนสารสนเทศ โดยใช้หลักการเว็บเชิงความหมาย
2. เพื่อศึกษาโครงสร้างของการตัดคำในภาษาไทย ก่อนที่จะนำไปเปรียบเทียบกับความสัมพันธ์ของกลุ่มคำที่สร้างขึ้นเพื่อใช้เป็นคำสำคัญ (Keyword) ในการแสดงผลการค้นคืนสารสนเทศ
3. เพื่อพัฒนาระบบค้นคืนสารสนเทศภาษาไทยให้แสดงผลลัพธ์ในการค้นคืนได้แม่นยำและตรงตามความต้องการของผู้ใช้ที่สุด

1.3 ขอบเขตของงานวิจัย

งานวิจัยนี้มีจุดมุ่งหมายที่จะศึกษาและพัฒนาระบบการตัดคำทำงานผ่านเว็บเบราว์เซอร์ เพื่อการค้นคืนสารสนเทศภาษาไทย โดยนำหลักการเชิงความหมายมาช่วยพัฒนาระบบ ซึ่งมุ่งเน้นไปที่การค้นคืนข้อมูลเทคโนโลยีสารสนเทศภาษาไทยโดยขอบเขตของงานวิจัยนี้จะประกอบด้วยการพัฒนาส่วนประกอบต่างๆ ของระบบบริการสืบค้นเชิงความหมายดังนี้

1. ศึกษาและพัฒนาระบบค้นคืนสารสนเทศโดยใช้หลักการค้นคืนเชิงความหมาย

2. ศึกษาแนวคิดเกี่ยวกับหลักการตัดคำภาษาไทย และการจัดกลุ่มความสัมพันธ์ของความหมายของคำในภาษาไทย เพื่อนำมาจัดการในส่วนของความสัมพันธ์ของคำเพื่อคัดกรองคำหลักในการค้นคืนได้ดีขึ้น

3. พัฒนาระบบค้นคืนสารสนเทศ โดยมีหลักการทำงานในรูปแบบของเว็บแอพพลิเคชันโดยมีขอบเขตของข้อมูลที่ใช้ในการทดสอบเป็นข้อมูลเฉพาะด้านเทคโนโลยีสารสนเทศ เพื่อนำมาใช้ในการตัดคำหลักในการค้นคืนและการตัดคำของเอกสารเพื่อนำมาสร้างเป็นดัชนีในการค้นคืน

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ทำให้การค้นคืนสารสนเทศโดยใช้คำหลักเป็นภาษาไทยได้ผลลัพธ์ที่เที่ยงตรงแม่นยำมากขึ้น สะดวกสำหรับผู้ใช้งานที่ต้องการค้นหาสารสนเทศที่เป็นภาษาไทย

2. ระบบสามารถคาดเดาคำหลักในการค้นคืนสารสนเทศเมื่อผู้ใช้งานพิมพ์ผิดได้ เหมาะสำหรับการค้นคืนด้วยคำหลักที่เขียนเลียนเสียงตามภาษาอังกฤษซึ่งมักมีการเขียนที่ไม่ถูกต้อง

3. สามารถนำไปพัฒนาต่อยอดโดยทำการพัฒนาการตัดคำในภาษาไทยเพื่อให้มีขอบเขตของข้อมูลที่ต้องการค้นหามากขึ้น

บทที่ 2
ปริทัศน์ วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้ได้นำเสนอแนวทางการนำการเทคนิคการตัดคำในภาษาไทย และเทคโนโลยีเชิงความหมายมาช่วยพัฒนาระบบค้นคืนสารสนเทศ ซึ่งผู้วิจัยได้ศึกษาทฤษฎี เทคโนโลยี และวรรณกรรมที่เกี่ยวข้อง เพื่อใช้เป็นแนวทางในการศึกษาและพัฒนาซึ่งประกอบไปด้วย ระบบบริการค้นคืนสารสนเทศ (Search Engine) เทคโนโลยีเชิงความหมาย (Semantic Technology) หลักการออนโทโลจี (Ontology) หลักการตัดคำในภาษาไทย (Thai Word Segmentation) ระบบฐานข้อมูลความสัมพันธ์ภาษาอังกฤษ WordNet งานวิจัยที่เกี่ยวข้องและบทสรุป

2.1 ระบบบริการค้นคืนสารสนเทศ (Search Engine)

สิทธิศักดิ์ บุญมาก (2006) ได้ให้คำจำกัดความว่า ระบบช่วยบริการค้นคืนสารสนเทศเป็นระบบซอฟต์แวร์ที่มีซอฟต์แวร์หุ่นยนต์ (Robot Software) ซึ่งบางครั้งเรียกว่า Spider หรือ Web Crawler ทำหน้าที่ท่องไปในเว็บไซต์ต่างๆ ในอินเทอร์เน็ตเพื่อรวบรวมเอกสารหรือข้อมูลต่างๆ บนเว็บไซต์ แล้วนำมาสร้างเป็นฐานดัชนี (Index) สำหรับค้นคืนเอกสารบนเว็บ เพื่อทำให้การสืบค้นกระทำได้อย่างรวดเร็ว ดังนั้นกระบวนการจัดเตรียมดัชนี จึงเป็นขั้นตอนสำคัญอย่างยิ่งสำหรับการค้นคืนสารสนเทศ ส่วนการค้นคืนสารสนเทศแบบออฟไลน์หรือการค้นคืนสารสนเทศในองค์กรต่างๆ ที่ไม่ได้ผ่านเครือข่ายอินเทอร์เน็ต การสร้างดัชนีมักถูกสร้างโดยการพัฒนาซอฟต์แวร์ขึ้นมาจัดการข้อมูลต่างๆ เพื่อสร้างเป็นดัชนี หรืออาจไม่มีการสร้างดัชนีสำหรับการค้นคืนสารสนเทศแต่พัฒนาระบบค้นคืนสารสนเทศให้ค้นหาโดยตรงจากฐานข้อมูลขององค์กร แต่มักจะมีปัญหาในเรื่องความเร็วของการค้นคืนสารสนเทศซึ่งจะทำการค้นคืนเอกสารได้ช้ากว่าการค้นคืนจากดัชนีในจำนวนข้อมูลที่เท่ากัน

เนื่องจากการเตรียมข้อมูลหรือดัชนีสำหรับการค้นคืน เป็นกระบวนการที่จัดเตรียมโดยใช้ซอฟต์แวร์ ทำให้สามารถปรับปรุงข้อมูลได้ตลอดเวลา โดยอาจจะถูกปรับปรุงโดยอัตโนมัติจากซอฟต์แวร์หรืออาจจะเป็นการปรับปรุงโดยผู้ดูแลระบบ จึงไม่มีปัญหาในกรณีที่มีข้อมูลสารสนเทศใหม่ๆ เข้ามาในระบบอินเทอร์เน็ต ตัวอย่างของผู้ให้บริการค้นคืนสารสนเทศที่มีอยู่ในปัจจุบันและเป็นที่นิยมเช่น Google Yahoo Bing เป็นต้น

ระบบค้นคืนสารสนเทศที่มีอยู่ในปัจจุบัน จะมีหลักการทำงานที่แตกต่างกันและการจัดอันดับผลลัพธ์ในการค้นคืนสารสนเทศที่แตกต่างกัน เพราะมีลักษณะการค้นคืนที่แตกต่างกันทำให้โดยทั่วๆ ไปแล้วจะมีการแบ่งออกเป็นหลายๆ ประเภทด้วยกัน แต่สามารถสรุปออกมาได้เป็น 3 ประเภทหลักๆ ดังนี้

2.1.1 Crawler Based Search Engine

ระบบค้นคืนสารสนเทศแบบ Crawler Based Search Engine นี้เป็นระบบค้นคืนบนอินเทอร์เน็ตแบบที่อาศัยการบันทึกข้อมูลและจัดเก็บข้อมูลเป็นหลัก ซึ่งส่วนใหญ่ระบบให้บริการค้นคืนสารสนเทศที่ได้รับความนิยมมักจะทำงานในรูปแบบนี้ เนื่องจากให้ผลการค้นคืนได้แม่นยำที่สุด และการประมวลผลการค้นคืนสารสนเทศสามารถทำได้รวดเร็วจึงทำให้มีบทบาทในการค้นคืนสารสนเทศมากที่สุด โดยระบบค้นคืนสารสนเทศประเภทนี้จะมีองค์ประกอบหลักๆ 2 องค์ประกอบด้วยกันคือ

ฐานข้อมูล โดยส่วนใหญ่แล้วระบบค้นคืนสารสนเทศประเภทนี้จะมีฐานข้อมูลเป็นของตัวเองที่มีระบบการประมวลผลและการจัดอันดับเฉพาะที่เป็นเอกลักษณ์ของตนเอง โดยฐานข้อมูลนี้จะเก็บข้อมูลรายละเอียดต่างๆ ที่ได้มาจากการเก็บข้อมูลโดยซอฟต์แวร์หุ่นยนต์ตามที่ผู้ให้บริการออกแบบมา

ซอฟต์แวร์ คือเครื่องมือหลักสำคัญที่สุดอีกส่วนหนึ่งสำหรับระบบค้นคืนสารสนเทศประเภทนี้เนื่องจากต้องอาศัยซอฟต์แวร์เล็กๆ ทำหน้าที่ตรวจหาและทำการจัดเก็บข้อมูล ที่อยู่บนเว็บหรือเอกสารต่างๆ ในรูปแบบของการทำสำเนาเหมือนกับต้นฉบับทุกอย่าง ซึ่งเราเรียกซอฟต์แวร์ชนิดนี้ว่าตัวแทนค้นหาบนเว็บไซต์ (Web Crawler หรือ Web Spider)

ตัวอย่างผู้ให้บริการที่ใช้รูปแบบการค้นคืนสารสนเทศลักษณะนี้ คือ www.google.com www.yahoo.com เป็นต้น

2.1.2 Web Directory หรือ Blog Directory

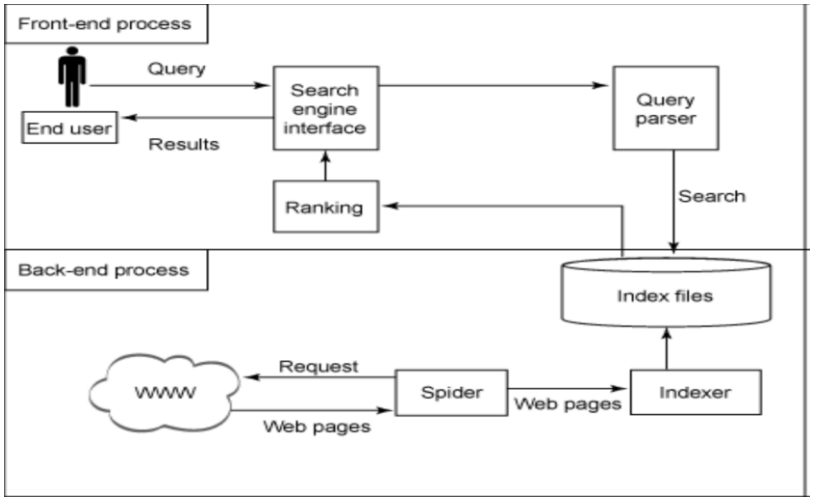
ระบบค้นคืนสารสนเทศแบบ Web Directory หรือ Blog Directory คือสารบัญเว็บไซต์ที่สามารถค้นหาข่าวสารข้อมูลด้วยหมวดหมู่ข่าวสารที่เกี่ยวข้องกัน ในปริมาณมากๆ คล้ายกับสมุดรายชื่อโทรศัพท์ ซึ่งจะมีการสร้างดัชนี มีการระบุหมวดหมู่อย่างชัดเจน ที่จะช่วยให้การค้นคืนสารสนเทศต่างๆ ตามหมวดหมู่นั้นๆ ได้รับการเปรียบเทียบอ้างอิง เพื่อหาข้อเท็จจริงได้ในขณะที่เรา

ค้นหาข้อมูล เพราะว่าเว็บไซต์ที่มีเนื้อหาคล้ายกันในหมวดหมู่เดียวกันนั้นมีมากมาย ทำให้เราเลือกที่จะค้นคืนข้อมูลสารสนเทศได้อย่างตรงประเด็นที่สุด และช่วยย่นระยะเวลาในการค้นคืนสารสนเทศได้ ตัวอย่างของผู้ให้บริการที่ใช้รูปแบบการค้นคืนสารสนเทศลักษณะนี้เช่น www.bloghints.com www.blogcatalog.com เป็นต้น

2.1.3 Meta Search Engine

ระบบค้นค้นคืนสารสนเทศแบบ Meta Search Engine คือการค้นคืนสารสนเทศที่ใช้หลักการค้นคืนโดยอาศัย Meta Tag ในภาษา HTML ซึ่งมีการประกาศชุดคำสั่งต่างๆ ซึ่งเป็นรูปแบบของ Text Editor ด้วยภาษา HTML เช่น ชื่อผู้พัฒนา คำค้นหา เจ้าของเว็บไซต์หรือคำอธิบาย ผลการค้นหาในส่วนนี้ไม่แม่นยำนัก เนื่องจากบางครั้งผู้ให้บริการหรือผู้ออกแบบเว็บไซต์นั้นๆ สามารถเลือกที่จะใส่ข้อมูลอะไรเข้าไปก็ได้เพื่อให้เกิดการถูกค้นคืนสารสนเทศของตนเองได้มากที่สุด หรือในอีกรูปแบบหนึ่งของการค้นคืนสารสนเทศลักษณะนี้คือ มีการอาศัยระบบบริการค้นคืนสารสนเทศอื่นๆหลายๆ แห่งมาประมวลผลรวมกัน โดยการส่งคำหลักที่ต้องการค้นหาของผู้ใช้งานไปให้ระบบค้นคืนสารสนเทศอื่นๆทำการค้นคืนแล้วจึงนำมาแสดงผลลัพธ์ ซึ่งทำให้ผลการค้นหาไม่เที่ยงตรงเท่าที่ควร ตัวอย่างของผู้ให้บริการที่ใช้รูปแบบการค้นคืนสารสนเทศลักษณะนี้เช่น www.metacrawler.com www.search.com เป็นต้น

จากที่กล่าวมาข้างต้นส่วนใหญ่รูปแบบการค้นคืนสารสนเทศมักมีรูปแบบที่คล้ายกันขึ้นอยู่กับการใช้งานแต่ละประเภท โดยอาจจะมองเป็นรูปแบบสถาปัตยกรรมโดยรวมได้ดังแสดงในรูปที่ 2.1



รูปที่ 2.1 แสดงสถาปัตยกรรมระบบสืบค้นสารสนเทศ

ที่มา : http://www.ibm.com/developerworks/web/library [8 October 2010]

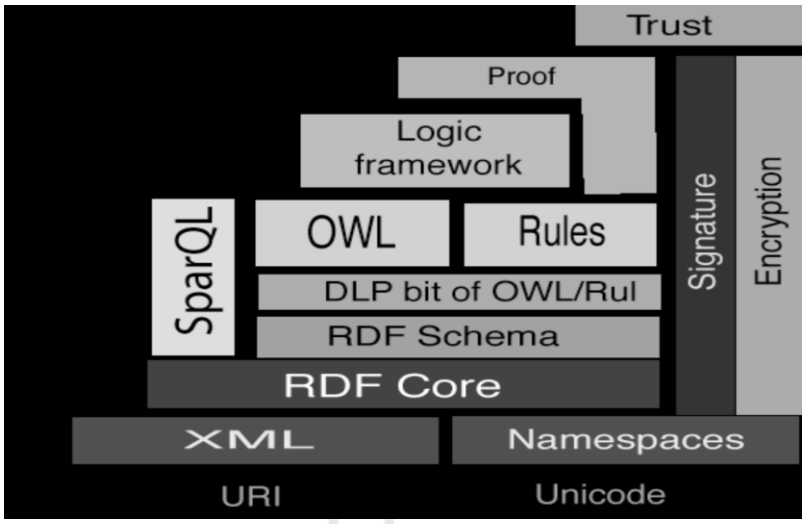
2.2 เทคโนโลยีเชิงความหมาย (Semantic Technology)

ในปัจจุบันข้อมูลที่ได้ถูกจัดทำขึ้นในเว็บ เป็นข้อมูลที่มีประโยชน์แต่ไม่เอื้อต่อเครื่องคอมพิวเตอร์ในการวิเคราะห์หรือการค้นคืนอย่างอัตโนมัติ เนื่องจากข้อมูลเหล่านั้นขาดโครงสร้างและเป็นเพียงชิ้นส่วนของเท็กซ์ (Text) ซึ่งมนุษย์สามารถเข้าใจความหมายแต่คอมพิวเตอร์ไม่สามารถเข้าใจความหมายได้

Tim Berners, James Hendler และ Ora Lassila (2001) ได้นำเสนอเทคโนโลยีเว็บเชิงความหมายหรือ Semantic Web Technology ซึ่งเป็นเทคโนโลยีที่ใช้ในการจัดเก็บและนำเสนอเนื้อหาแบบมีโครงสร้างรวมถึงสามารถที่จะวิเคราะห์จำแนกหรือจัดแบ่งได้ว่าข้อมูลที่ปรากฏนั้นมีความสัมพันธ์กับข้อมูลอื่นๆ ในแต่ละระดับอย่างไร กล่าวคือเป็นการจัดเก็บและนำเสนอโดยมีความสัมพันธ์แบบลำดับขั้น (Hierarchy) นั่นเอง

ประเด็นหลักที่ทำให้เกิดการพัฒนาเทคโนโลยีเว็บเชิงความหมายก็คือ สาเหตุจากการที่เว็บไซต์ในปัจจุบันที่ส่วนใหญ่ถูกเรียกว่าเป็น Syntactic หรือ Hypermedia Web มีปัญหาในเรื่องของ Information Overload เพราะว่าข้อมูลที่ค้นคืนมานั้นผลลัพธ์ที่ได้ไม่มีประสิทธิภาพเพียงพอและไม่สะดวกในการที่จะนำไปใช้งานต่อเพราะการค้นหาด้วย Keyword ทั่วๆ ไปนั้นเครื่องคอมพิวเตอร์ไม่สามารถทำความเข้าใจและประมวลความหมายหรือความสัมพันธ์ของคำนั้นๆ ได้อย่างตรงประเด็น ผลของการค้นคืนที่ได้กลับมาจึงเป็นการแสดงผลลัพธ์ทุกๆ เรื่องที่มีคำๆ นั้น และสร้าง Hyperlink เพื่อให้เชื่อมโยงไปยังข้อมูล โดยไม่รู้ว่านั่นคือคำที่อยู่ในเรื่องที่ผู้ใช้งานต้องการหรือไม่

แนวทางของ Semantic Web Technology ที่ช่วยแก้ปัญหาดังกล่าวก็คือ Semantic Web Technology มีการ Provide Common Framework ซึ่งทำให้ข้อมูลสามารถ Share และ Reused ข้าม Application หรือ Community ที่มีการระบุขอบเขตได้โดยที่เครื่องคอมพิวเตอร์สามารถเข้าใจองค์ประกอบของข้อมูลซึ่งมีการแนบ Domain Theory เช่น รูปแบบของการอ้างอิง Class แม่ของข้อมูล รูปแบบนี้อาจเรียกว่าเป็น Ontology ซึ่งสามารถบอกระดับความสัมพันธ์ของข้อมูลได้



รูปที่ 2.2 แสดงสถาปัตยกรรมของเว็บเชิงความหมาย

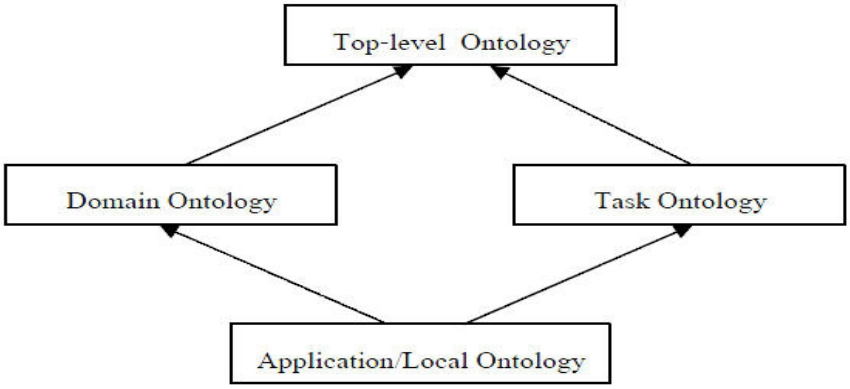
ที่มา : http://www.semanticfocus.com [10 October 2010]

2.3 หลักการออนโทโลจี (Ontology)

Antonio G. และ Van Harmelen F. (2004) ได้ให้ความหมายเกี่ยวกับหลักการออนโทโลจีไว้ว่า ออนโทโลจีถูกสร้างขึ้นมาเพื่อใช้บรรยายแนวคิดของโดเมนหรือขอบเขตความสนใจใดๆ ในรูปของสิ่งต่างๆ ที่อยู่ภายใต้หรือภายในโดเมนและความสัมพันธ์ระหว่างสิ่งเหล่านั้นโดยแสดงออกมาในรูปแบบคำสามัญ (Common Word) เพื่อให้สามารถเข้าใจและนำไปใช้ร่วมกันได้ ในภาษาธรรมชาติ (Natural Language) สามารถแสดงโดยใช้คำศัพท์ (Vocabulary) และประโยค (Sentence) ซึ่งเกิดจากการรวมคำศัพท์ต่างๆ เข้าไว้ด้วยกันเพื่อแสดงความสัมพันธ์ระหว่างคำศัพท์เหล่านั้น ส่วนการนำไปใช้ด้านคอมพิวเตอร์จะแสดงในรูปแบบของระบบสัญลักษณ์ (Notation) เช่น คลาส (Class) อินสแตนซ์ (Instance) ความสัมพันธ์ (Relation) คุณสมบัติ (Property) กฎ (Rule) เป็นต้น โดยใช้ภาษาสำหรับแสดงความรู้ (Knowledge Representation Language) ซึ่งมีความชัดเจนและเที่ยงตรงมากกว่าคำศัพท์และประโยคในภาษาธรรมชาติ ทั้งนี้เพื่อให้ซอฟต์แวร์และเครื่องมือสามารถนำไปประมวลผลได้

2.3.1 ประเภทของออนโทโลจี

รูปที่ 2.3 ประเภทของออนโทโลจี จะประกอบไปด้วย Top-Level Ontology, Domain Ontology, Task Ontology และ Application Ontology โดยมีรายละเอียดดังนี้



รูปที่ 2.3 ประเภทของออนโทโลจี

ที่มา : http://mrkrich.blogspot.com/2009/10/ontology.html [10 October 2010]

1. ออนโทโลจีระดับบน (Top-Level Ontology) เป็นออนโทโลจีที่ประกอบไปด้วยเบสคลาส (Based Class) และกำหนดคุณสมบัติเพื่ออธิบายคลาสหรือกำหนดความสัมพันธ์ระหว่างคลาส โดยสามารถนำไปใช้งานจริงได้ในโดเมนทั่วไป (Generic Domain)

2. ออนโทโลจีสำหรับกิจกรรม (Task Ontology) เป็นออนโทโลจีที่พัฒนาขึ้นเพื่อตอบสนองการทำงานของกิจกรรมต่างๆ โดยอาศัยการถ่ายทอดคุณลักษณะเฉพาะของกิจกรรมจากออนโทโลจีระดับบน

3. ออนโทโลจีระดับโดเมน (Domain Ontology) เป็นออนโทโลจีที่ตอบสนองต่อโดเมน โดยอาศัยการถ่ายทอดคุณลักษณะเฉพาะของโดเมนจากออนโทโลจีระดับบน

4. ออนโทโลจีระดับแอพพลิเคชัน (Application Ontology) เป็นออนโทโลจีที่ถูกจำกัดการใช้งานในโดเมนที่มีความจำเพาะเจาะจง (Specific Domain) โดยอาศัยการถ่ายทอดคุณลักษณะเฉพาะของโดเมนจากออนโทโลจีสำหรับโดเมนหรือออนโทโลจีสำหรับกิจกรรมก็ได้

2.3.2 การสร้างและการพัฒนาออนโทโลจี

การสร้างและการพัฒนาออนโทโลจี (Create Ontology) เป็นงานที่ต้องอาศัยความรู้ความเข้าใจในความสัมพันธ์ของสิ่งต่างๆ ในโดเมนเป็นอย่างดี โดยผู้ที่ทำหน้าที่สร้างออนโทโลจีนั้นคือผู้เชี่ยวชาญโดเมน (Domain Expert) ซึ่งมีขั้นตอนการสร้างออนโทโลจีดังนี้

ขั้นตอนการกำหนดขอบเขตของการพัฒนา (Define Scope) ในขั้นตอนนี้จะกำหนดความต้องการของระบบที่จะนำออนโทโลจีไปใช้งาน โดยทำการระบุขอบเขตและวัตถุประสงค์ของการพัฒนา ซึ่งในการระบุความต้องการจะมีผลกระทบต่อการออกแบบ การประเมินผลและการนำกลับมาใช้ใหม่ของออนโทโลจี

ขั้นตอนการพิจารณาถึงการนำกลับมาใช้ใหม่ของออนโทโลจี (Reuse Ontology) ในการนำออนโทโทโลจีที่มีอยู่กลับมาใช้ใหม่จะช่วยลดความลำบากในการพัฒนาลง ซึ่งในการนำกลับมาใช้ใหม่นั้นจะต้องให้ความสำคัญกับ Ontology Library Systems (OLS) เนื่องจากเป็นเครื่องมือที่ช่วยในการจัดกลุ่ม (Grouping) การรวบรวม (Integration) การดูแลรักษา (Maintenance) การนำส่ง (Mapping) และการออกเวอร์ชัน (Versioning)

ขั้นตอนการระบุรายละเอียดเงื่อนไขที่เกี่ยวข้อง (Define Relation) โดยขั้นตอนนี้จะทำการระบุรายละเอียดและเงื่อนไขที่เกี่ยวข้อง พิจารณาถึงรายละเอียดของเทอมทั้งหมดว่าต้องการสื่อถึงเรื่องใด มีคุณสมบัติอย่างไร โดยไม่คำนึงถึงความซ้ำซ้อนกันของคอนเซพท์และคุณสมบัติต่างๆ

ขั้นตอนการกำหนดคลาส (Define Classes) โดยคลาส หมายถึงคอนเซพท์ที่อยู่ในโดเมนที่ประกอบด้วยส่วนประกอบต่างๆ ซึ่งในการพัฒนาลำดับของคลาสมีอยู่ด้วยกันหลายวิธี โดยวิธีการพัฒนาลำดับของคลาสที่นิยมได้แก่

1. วิธีแบบบนลงล่าง (Top-Down) คือกระบวนการพัฒนาที่เริ่มจากการกำหนดนิยามของคลาสทั้งหมดในโดเมนและขอบเขตของคลาส
2. วิธีแบบล่างขึ้นบน (Bottom-Up) คือกระบวนการที่เริ่มจากการกำหนดโดยระบุคลาส คลาสจะถูกแยกกลุ่มออกมาก่อนจะถูกนำไปใส่ในคลาสแม่
3. วิธีแบบผสม (Combination) คือการรวมวิธีบนลงล่างและล่างขึ้นบนเข้าด้วยกันและกำหนดคลาสขึ้นมาก่อนโดยกำหนดคลาสขึ้นมาก่อนและวางหลักกว้างๆไว้ก่อนจะระบุว่าอย่างไหนเหมาะสม

ขั้นตอนการกำหนดคุณสมบัติของคลาส (Define Properties) ซึ่งจะต้องกำหนดประเภทให้กับคุณสมบัติของคลาสด้วย และต้องพิจารณาว่าคลาสมีคุณสมบัติแบบง่าย (Simple Properties)

เช่นมีค่าดั้งเดิมเป็นข้อความหรือตัวเลข เป็นต้น หรือมีคุณสมบัติที่ซับซ้อน (Complex Properties) เช่น วัตถุต่างๆ ที่เป็นอินสแตนซ์ เป็นต้น

ขั้นตอนการกำหนดเงื่อนไขให้กับคุณสมบัติ (Define Properties) หรือกำหนดฟาร์เซ็ทให้กับสล็อท ซึ่งสล็อทจะมีฟาร์เซ็ทที่แตกต่างกัน โดยจะต้องพิจารณาถึง Slot Cardinality โดเมนและเรนจ์ของสล็อท

ขั้นตอนการสร้างอินสแตนซ์ของคลาส (Create Instance) โดยจะต้องทำการกำหนด Individual Instance ของคลาสให้กับสล็อทสำหรับอินสแตนซ์ของเฟรมด้วย

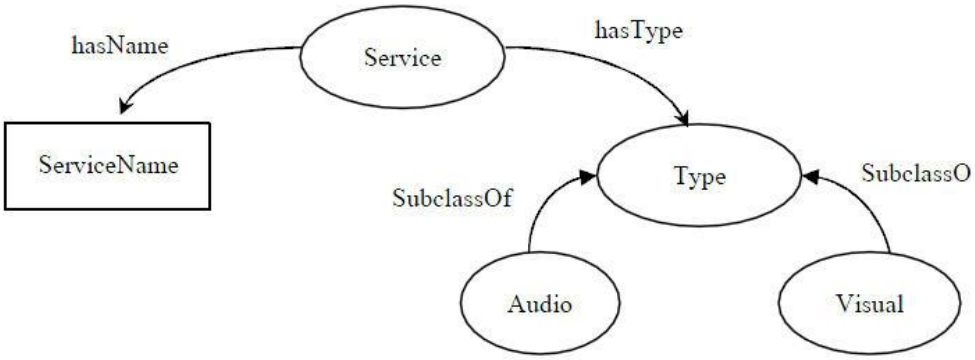
2.3.3 ส่วนประกอบหลักของออนโทโลจี

การกำหนดข้อมูลเค้าร่างสำหรับอธิบายข้อมูลเชิงความหมายคือ การกำหนดออนโทโลจีระดับบนเพื่อเป็นโมเดลแสดงโครงสร้างการอธิบายข้อมูลเชิงความหมาย ซึ่งจะต้องกำหนดคลาส (Class) พร็อพเพอร์ตี้ (Property) และเงื่อนไข (Restriction) สำหรับการอธิบายข้อมูลเค้าโครงร่างข้อมูลแม่แบบสำหรับการอธิบายข้อมูลเชิงความหมาย ซึ่งในการอธิบายข้อมูลประกอบด้วย

คลาส (Class) หรือคอนเซ็ปต์ (Concept) หรือ (Category) เป็นตัวแสดงถึงความรู้ที่เราสนใจและอธิบายได้ว่าคลาสต่างๆ บรรจุอะไรไว้ในโดเมน ซึ่งคลาสเป็นส่วนที่จะต้องพิจารณาอย่างละเอียดในการพัฒนาออนโทโลจี

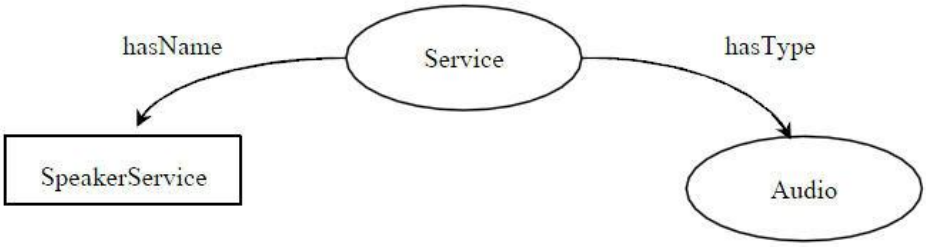
พร็อพเพอร์ตี้ (Property) หรือความสัมพันธ์ (Relation) หรือสล็อท (Slot) แสดงถึงการกำหนดความสัมพันธ์ (Relation) หรือคุณลักษณะของคลาส เพื่อเชื่อมโยงระหว่างคลาสด้วยการระบุพร็อพเพอร์ตี้ที่สามารถกำหนดได้ด้วยการประกาศให้เป็นค่าคงที่

จากรูปที่ 2.4 เป็นรูปที่แสดงตัวอย่างข้อมูลเค้าร่างที่อธิบายข้อมูลรายละเอียดของคลาสเซอร์วิสโดยมีการประกาศคลาสและพร็อพเพอร์ตี้ได้แก่ คลาส Service ซึ่งมีความสัมพันธ์ hasType กับคลาส Type ที่มี SubclassOf สองคลาสคือคลาส Audio และคลาส Visual นอกจากมีความสัมพันธ์กับคลาส Type แล้วคลาส Service ยังมีความสัมพันธ์ hasName กับค่า ServiceName อีกด้วย



รูปที่ 2.4 ตัวอย่างข้อมูลเค้าร่างอธิบายข้อมูลรายละเอียดคลาสเซอร์วิส
ที่มา : David Martin (2004) http://www.w3c.org [10 October 2010]

ข้อมูลอินสแตนซ์ (Instance Data) คือการอธิบายรายละเอียดของข้อมูลซึ่งใช้ข้อมูลเค้าร่างเป็นแม่แบบในการอธิบาย ดังรูปที่ 2.5 แสดงตัวอย่างข้อมูลอินสแตนซ์อธิบายรายละเอียดเซอร์วิส



รูปที่ 2.5 ตัวอย่างข้อมูลเชิงอินสแตนซ์อธิบายรายละเอียดเซอร์วิส
ที่มา : David Martin (2004) http://www.w3c.org [10 October 2010]

การอธิบายข้อมูลเชิงความหมายที่นำไปใช้งานในโปรแกรมประยุกต์ อาจจะเลือกอธิบายข้อมูลเค้าร่างเพียงอย่างเดียวก็ได้ ทั้งนี้ขึ้นอยู่กับความจำเป็นในการออกแบบและการใช้งานออนโทโลจีในโปรแกรมประยุกต์

2.3.4 เครื่องมือที่ใช้สำหรับการพัฒนาออนโทโลจี

เครื่องมือที่ใช้สำหรับการพัฒนาออนโทโลจี มีความสำคัญต่อกระบวนการพัฒนาออนโทโลจีในทุกขั้นตอน ตั้งแต่การสร้าง การจัดเก็บ การจัดการความรู้ การดูแลรักษา จนกระทั่งการประเมินผล

Knublauch H., Ferguson W., Natalya F. และ Musen A. (2004) ได้เสนอเครื่องมือเกี่ยวกับการพัฒนาออนโทโลจีดังนี้ Protégé เป็นซอฟต์แวร์สำหรับการสร้างออนโทโลจีและระบบฐานความรู้ (Knowledge Based System) ซึ่งสนับสนุนภาษา OWL (Web Ontology Language) ผ่าน OWL Plug-in มีส่วนการติดต่อผู้ใช้งานแบบกราฟิก (Graphical User Interface : GUI) ที่รองรับการทำงานแบบหลายผู้ใช้งาน (Multi Users) สามารถจัดเก็บออนโทโลจีในรูปแบบแฟ้มข้อมูลและฐานข้อมูลเชิงสัมพันธ์ มีเครื่องมือสำหรับการสร้างโดเมนของออนโทโลจี และรูปแบบข้อมูลที่สะดวกในการป้อนข้อมูล โดยอนุญาตให้ผู้ใช้งานสามารถทำงานพร้อมกันบนคลาสหรืออินสแตนซ์เดียวกัน ซึ่งผู้ใช้สามารถนำโดเมนของออนโทโลจีกลับมาใช้ใหม่ และช่วยแก้ปัญหาที่เกิดขึ้นในการพัฒนาส่วนของวิธีการ โดยหลายๆ แอพพลิเคชั่นสามารถใช้งานโดเมนเดียวกัน เพื่อแก้ปัญหาที่แตกต่างกัน และวิธีการนั้นสามารถนำไปประยุกต์ใช้กับออนโทโลจีที่ต่างกันได้ใน OWL Plug-in มีคุณสมบัติหลักๆ ได้แก่ การบรรจุ (Load) และการบันทึก (Save) ออนโทโลจีในรูปแบบ OWL (Ontology Web Language) และ RDF (Resource Description Framework) การตรวจแก้ (Edit) และแสดงผลคลาสรวมถึงคุณสมบัติของภาษา OWL ในรูปแบบกราฟิก ความสามารถในการกำหนดลักษณะเชิงตรรกะของคลาส (Logical Class Characteristic) ได้ในรูปแบบนิพจน์ (Expression) ความสามารถในการประมวลผลร่วมกับเครื่องมืออนุมาน (Inference Engine) อื่นๆได้ และความสามารถในการตรวจแก้อินสแตนซ์ของภาษา OWL เพื่อใช้อธิบายเว็บเชิงความหมาย

นอกจากการใช้ Protégé เพื่อสร้างออนโทโลจีแล้ว ในการสร้างโปรแกรมประยุกต์ด้านเว็บเชิงความหมายแล้วยังสามารถใช้ Jena ซึ่ง McBride B. (2002) ได้ให้คำจำกัดความว่า Jena เป็นเอพีไอภาษาจาวา (Java API) จาก W3C (World Wide Web Consortium) ใช้สำหรับสร้างโปรแกรมสนับสนุนสภาพแวดล้อมในการพัฒนาโปรแกรมสำหรับจัดการเอกสาร RDF RDFS (Resource Description Framework Schema) และ OWL รวมถึงการอนุมานโดยใช้กฎ (Rule-Based Inference Engine) คุณสมบัติของจีนาซึ่งมี OWL API และ RDF API สามารถนำเข้าและส่งออกอาร์ดีเอฟโมเดลในรูปแบบ อาร์ดีเอฟเอ็กซ์เอ็มแอล โนเทชัน 3 และเอ็นทริปเปิล มีการสนับสนุนการจัดการ RDF Model ในหน่วยความจำและจากรากฐานข้อมูล และสนับสนุนภาษา RDQL ซึ่งเป็นภาษา

สำหรับสืบค้น (Query Language) สำหรับอาร์ดีเอฟโมเดลที่อยู่ในสภาพแวดล้อมของจีนา โดยมอง RDF Model เป็นข้อมูลแบบ Triple

ภาษาที่ใช้สำหรับอธิบายออนโทโลจีในงานวิจัยนี้จะประกอบไปด้วยภาษา RDF ภาษา RDFS ภาษา OWL (Ontology Web Language) และภาษา OWL-S (OWL-Based Web Service Ontology) ซึ่งมีรายละเอียดดังนี้

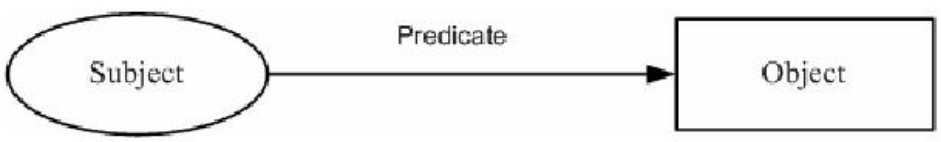
2.3.4.1 RDF (Resource Description Framework)

Singh M. และ Huhns M. (2005) ได้ให้ความหมายไว้ว่า RDF เป็นภาษามาตรฐานโดย W3C Recommendation สำหรับอธิบายทรัพยากรและความสัมพันธ์ของทรัพยากรโดยอาศัยแบบจำลองพื้นฐานรูปแบบอิเล็กทรอนิกส์ เช่น แฟ้มข้อมูล (File) หรือแนวคิดเกี่ยวกับสิ่งต่างๆ เช่น คน สัตว์ โดยลักษณะแบบจำลองในการแทนข้อมูลด้วย RDF Model หรือ RDF Statement ซึ่งจะอธิบายทรัพยากรหนึ่งๆ ในรูปของ Triple ที่ประกอบไปด้วยองค์ประกอบสามส่วนดังรูปที่ 2.6 ได้แก่

Subject แสดงถึงทรัพยากรซึ่งถูกบรรยายด้วย Predicate และ Object Predicate แสดงถึงส่วนที่อธิบายคุณสมบัติของทรัพยากรซึ่งแสดงความสัมพันธ์ระหว่าง Subject และ Object

Object แสดงถึงส่วนที่ถูกอ้างถึงซึ่งมีความสัมพันธ์กับ Subject ด้วย Predicate Object สามารถเป็นอย่างใดอย่างหนึ่งระหว่างทรัพยากร (Source) และค่าสัญพจน์ (Literal)

จากรูปที่ 2.6 แสดงการแทนที่ข้อมูลด้วยกราฟซึ่งประกอบด้วยจุดยอดและด้านของกราฟ โดยจุดยอดเป็นโหนดข้อมูลที่เรียกว่า Subject ใช้แสดงค่าวัตถุหรือทรัพยากร Object จะถูกแทนที่ด้วยโหนดข้อมูลใช้แสดงคุณสมบัติของทรัพยากรหรือคำอธิบายต่างๆ โดยมี Predicate แสดงความสัมพันธ์ระหว่าง Subject และ Object ทั้ง Subject, Predicate และ Object ซึ่งถูกระบุด้วย URI (Uniform Resource Identifier) ยกเว้น Object ที่อยู่ในรูปแบบ Literal ซึ่งประกอบด้วย 2 ส่วนคือ Namespace และ Local Name อย่างไรก็ตามโครงสร้างของภาษา RDF นั้นไม่สามารถนิยามการสร้างออนโทโลจีได้ทั้งหมดเป็นเพียงพื้นฐานสำหรับนิยามออนโทโลจีเบื้องต้นเท่านั้น



รูปที่ 2.6 ส่วนประกอบของ RDF โมเดล

ที่มา : http://www.altova.com/semantic_web.html [11 October 2010]

2.3.4.2 RDFS (Resource Description Framework Schema)

Brickley D. และ Guha R.V. (2004) ได้ให้ความหมายของ RDFS ว่าเป็นภาษามาตรฐานที่พัฒนาโดย W3C Recommendation ใช้โครงสร้างพื้นฐานของภาษา RDF เพื่อบรรยายลำดับชั้นของแนวคิดและคุณสมบัติ (Hierarchy of Concept and Properties) เพื่อสร้างออนโทโลจีอย่างง่ายทำให้สามารถอนุมาน (Inference) หาข้อเท็จจริง (Fact) นอกเหนือจากที่ประกาศให้ทรัพยากรได้ โดย RDSF ประกอบด้วยอีลีเมนต์หลักดังต่อไปนี้

<rdfs:Class> ได้แก่ อีลีเมนต์สำหรับอธิบายคลาส

<rdfs:ClassOf> ได้แก่ อีลีเมนต์สำหรับอธิบายคุณสมบัติของคลาสประกอบด้วยอีลีเมนต์ย่อย คือ <rdfs:domain> และ <rdfs:range>

<rdfs:domain> ได้แก่ อีลีเมนต์สำหรับกำหนดคลาสของ Subject

<rdfs:range> ได้แก่ อีลีเมนต์สำหรับกำหนดคลาสของ Object

ด้วยข้อจำกัดของคุณสมบัติของคลาสใน RDFS ทำให้ไม่เพียงพอต่อการบรรยายข้อมูลให้มีความหมายเชิงตรรกะได้ ดังนั้นจึงได้มีการกำหนดภาษาใหม่ที่ใช้บรรยายโครงสร้างความสัมพันธ์และความหมายเชิงตรรกะคือภาษา OWL

2.3.4.3 OWL (Ontology Web Language)

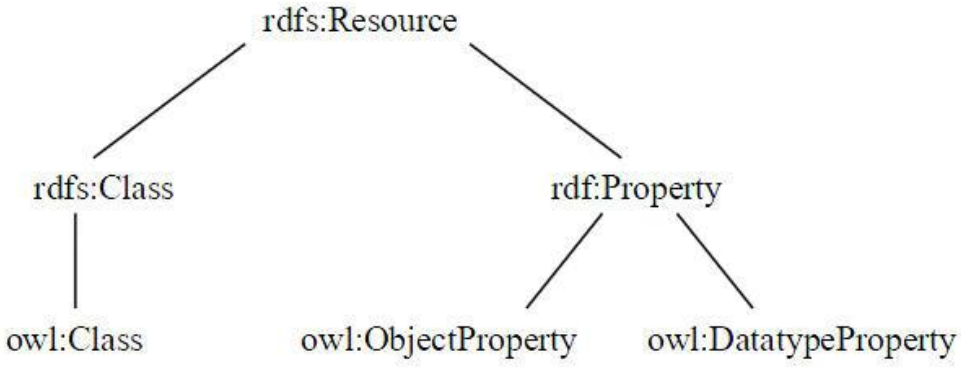
Mark W. (1995) ได้อธิบายว่า OWL เป็นภาษาออนโทโลจีมาตรฐานจาก W3C Recommendation ซึ่งใช้โครงสร้างพื้นฐานของภาษา RDFS และเป็นส่วนหนึ่งของกลุ่มภาษาที่ W3C กำหนดให้ใช้กับเว็บเชิงความหมาย เพื่อให้สามารถอธิบายความหมายของเอกสารได้มากขึ้น โดย OWL เป็นภาษาที่พัฒนามาจาก DAML+OIL (DARPA Agent Markup Language + Ontology Inference Layer) สนับสนุนคำศัพท์ที่ใช้อธิบายคลาสและคุณสมบัติเพิ่มเติมจากภาษา RDFS ภาษา OWL สามารถแบ่งออกได้เป็น 3 ภาษาย่อยคือ

1. OWL Lite เป็นภาษาย่อยขั้นต้นของ OWL ออกแบบมาเพื่อสนับสนุนการใช้งานเบื้องต้น สนับสนุนการจัดแบ่งลำดับชั้นของคลาส และการกำหนดข้อบังคับของภาษา OWL เบื้องต้น ซึ่งถูกออกแบบมาให้ง่ายในการพัฒนาและมีการเตรียมฟังก์ชันในการใช้งานต่างๆ สำหรับเริ่มใช้งานในการเขียนภาษา OWL

2. OWL DL เป็นภาษาย่อยขั้นต่อมาของ OWL สามารถอธิบายออนโทโลจีได้ละเอียดกว่า OWL Lite จัดให้มีคุณสมบัติที่เหมาะสมด้านการใช้งานฐานข้อมูล และการแทนที่ความรู้ที่ตั้งอยู่บนพื้นฐานของการอธิบายด้วยเหตุผลทางตรรกะ และสามารถกำหนดเวลาที่แน่นอนในการประมวลผลได้

3. OWL Full เป็นภาษาย่อยขั้นสูงของภาษา OWL ออกแบบมาเพื่อให้สนับสนุนผู้ใช้งานที่ต้องการความครบถ้วนและมีโครงสร้างภาษาที่สมบูรณ์แบบ สามารถอธิบายออนโทโลจีได้ละเอียดที่สุด แต่ไม่สามารถกำหนดเวลาที่แน่นอนในการประมวลผลได้

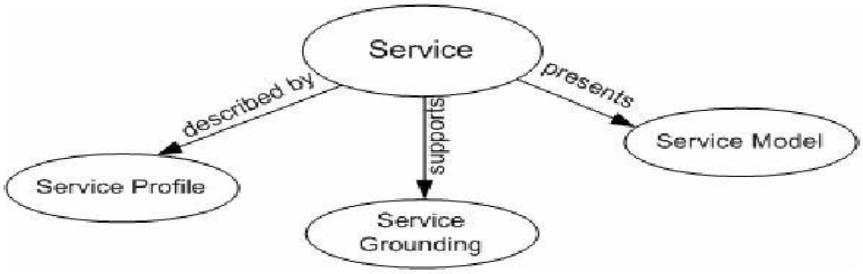
เนื่องจาก OWL เป็นภาษาที่ถูกขยายมาจากภาษา RDF ดังนั้นการบรรยายข้อมูลออนโทโลจีด้วย OWL จึงบรรยายข้อมูลด้วยโครงสร้างภาษา RDF หรือการแทนค่าข้อมูลในรูปแบบของ RDF Triple ข้อมูลในภาษา OWL จึงมีการบรรยายข้อมูลผสมกันระหว่างการใช้ RDF/RDFS และ XML Syntax ซึ่งแบ่งตามประเภทของการใช้งาน โดยในรูปที่ 2.7 จะแสดงถึงความสัมพันธ์ระหว่างคลาสของ OWL และ RDF/RDFS



รูปที่ 2.7 แสดงความสัมพันธ์ระหว่างคลาสของ OWL และ RDF/RDFs
ที่มา : David Martin. (2004). http://www.w3c.org. [10 October 2010]

2.3.4.4 OWL-S (OWL-Based Web Service Ontology)

Martin D., et al. (2004) ได้อธิบายเกี่ยวกับ OWL-S ว่าเดิมคือ DAML-S เป็นภาษาออนโทโลจีสำหรับเว็บเชิงความหมายที่พัฒนาต่อมาจากภาษา OWL ซึ่งใช้สำหรับอธิบายคุณสมบัติและความสามารถของเว็บเซอร์วิส โดยจะแบ่งรายละเอียดออกเป็นสามส่วนคือ เซอร์วิสโพรไฟล์ (Service Profile) เซอร์วิสโมเดล (Service Model) และเซอร์วิสกราวน์ดิง (Service Grounding) ดังแสดงในรูปที่ 2.8



รูปที่ 2.8 โครงสร้างของ OWL-S

ที่มา : David Martin (2004) http://www.w3c.org [10 October 2010]

Service Profile จะอธิบายรายละเอียดทั่วไปของเว็บเซอร์วิสที่เกี่ยวข้องกับการประกาศและการค้นหาบริการ โดย Service Profile จะแบ่งคุณสมบัติออกเป็นสองกลุ่มคือคุณสมบัติเชิงฟังก์ชัน (Function Properties) เช่น ข้อมูลเข้า (Input) ข้อมูลออก (Output) เป็นต้น เงื่อนไขก่อนการทำงาน (Preconditions) และผลกระทบ (Effect) ซึ่งข้อมูลเหล่านี้จะสืบทอดมาจาก Service Model คุณสมบัติที่ไม่เป็นฟังก์ชัน (Nonfunctional Properties) เช่น ชื่อบริการ (Service Name) สารสนเทศที่เกี่ยวกับผู้ให้บริการ (Provider Information) ประเภทของบริการ (Service Category) เป็นต้น

Service Model จะอธิบายว่าบริการนี้ทำงานอย่างไร โดยจะอธิบายลักษณะการทำงานของบริการในรูปแบบพฤติกรรมเชิงฟังก์ชัน (Functional Behavior) กระบวนการภายในของการบริการ (Internal Processes of The Service) และกระแสงาน (Workflow)

Service Grounding จะอธิบายถึงรายละเอียดของการติดต่อกับบริการ โดยจะมีการจับคู่ (Mapping) ระหว่างข้อกำหนดเชิงนามธรรมของ OWL-S และข้อกำหนดเชิงรูปธรรมของเว็บเซอร์วิส เช่น WSDL เป็นต้น

2.4 หลักการตัดคำในภาษาไทย

การตัดคำภาษาไทย (Thai Words Segmentation) ได้รับการพัฒนาขึ้นมาโดยใช้วิธีการต่างๆ กัน เนื่องจากการตัดคำเป็นกระบวนการพื้นฐานของการประมวลผลภาษาธรรมชาติเช่น การวิเคราะห์เสียงพูด ในการตัดคำภาษาไทยนั้นได้มีผู้คิดค้นวิธีที่จะแยกคำแต่ละคำออกจากประโยคซึ่งมีการเขียนติดกันไปอย่างต่อเนื่องทั้งประโยค ในหัวข้อนี้ผู้วิจัยจะกล่าวถึงวิธีและเทคนิคการตัดคำโดยอาศัยอักขรวิธี เป็นหลักการพื้นฐานการประสมคำ ซึ่งมีรายละเอียดดังต่อไปนี้

2.4.1 วิธีที่ใช้ตัดคำ

วิรัช ศรเลิศล้ำวาณิช และคณะ (2536) ได้เสนอวิธีการตัดคำโดยแบ่งหลักการตัดคำออกเป็น 3 ประเภทใหญ่ๆ ได้แก่ การใช้กฎ การใช้พจนานุกรมและการใช้คลังข้อความ

1. การใช้กฎ การตัดคำโดยการใช้กฎเป็นการตรวจสอบกฎเกณฑ์ทางอักขระวิธีที่กำหนดลักษณะการประสมอักษร ลักษณะการเว้นวรรค และการขึ้นย่อหน้า เพื่อใช้เป็นเกณฑ์ในการกำหนดขอบเขตของคำ วิธีการนี้จะมีข้อจำกัดในการทำงาน คือ ความถูกต้องของการตัดคำในระดับพยางค์สูงแต่ความถูกต้องของการตัดคำค่อนข้างต่ำ แต่ข้อดีของวิธีนี้คือมีความรวดเร็วในการทำงานและใช้ทรัพยากรน้อย

2. การใช้พจนานุกรม การตัดคำโดยพจนานุกรมเป็นการตัดคำโดยใช้สายอักขระ (String) มาเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรม ซึ่งวิธีนี้จะต้องทำการจัดเก็บคำไว้ในพจนานุกรม โดยการจัดเก็บคำไว้ในพจนานุกรมนี้ต้องดำเนินการโดยผู้เชียวชาญด้านภาษา เนื่องจากถ้าหากมีการผิดพลาดของการเขียนคำจะทำให้ระบบนำคำที่ผิดไปใช้งาน ดังนั้นจึงต้องทำการตรวจสอบคำให้ถูกต้องก่อนจัดเก็บลงพจนานุกรม วิธีนี้ทำให้ได้ความถูกต้องในการตัดคำสูงกว่าการใช้กฎแต่จะใช้เวลามากกว่าการใช้กฎ ซึ่งความเร็วขึ้นกับจำนวนคำที่มีอยู่ในพจนานุกรมด้วย

3. การใช้คลังข้อมูล การตัดคำโดยใช้คลังข้อมูล เป็นการตัดคำโดยนำวิธีการทางสถิติเข้ามาใช้ในการประมวลภาษาโดยใช้คลังข้อมูลทางภาษาเป็นฐานความรู้เก็บค่าความถี่ที่ใช้ในการตัดคำ ซึ่งการตัดคำโดยใช้คลังข้อมูลแบ่งออกเป็น 2 วิธี คือการตัดคำโดยอาศัยความน่าจะเป็น (Probabilistic Word Segmentation) และวิธีการตัดคำโดยอาศัยคุณลักษณะของคำ (Feature-Based Word Segmentation) วิธีการตัดคำโดยอาศัยค่าความน่าจะเป็นจะเป็นการตัดคำโดยใช้แบบจำลองเอนแกรม (Word N-Gram Model) ในการหารูปแบบของการตัดคำและลำดับคำที่เป็นไปได้มากที่สุด โดยวิธีการนี้จะต้องมีการใช้คลังข้อมูลที่มีการตัดคำและกำกับหมวดคำที่เตรียมเอาไว้แล้ว ซึ่งวิธีการนี้ผลลัพธ์ที่ได้จะเป็นการเลือกรูปแบบการตัดคำที่มีความน่าจะเป็นมากที่สุดวิธีการตัดคำโดยอาศัยคุณลักษณะของคำ จะเป็นการแก้ข้อผิดพลาดของการตัดคำโดยอาศัยค่าความน่าจะเป็นของ

การจำกัดหมวดคำที่จะเป็นแบบจำลองในการตัดคำ ซึ่งวิธีการตัดคำโดยอาศัยคุณลักษณะของคำจะเป็นวิธีการแบบผสม (Hybrid Approach)

2.4.2 เทคนิคที่ช่วยในการตัดคำ

วิรัช ศรเลิศล้ำวาณิช และคณะ (2536) ได้นำเสนอเทคนิคที่ใช้ในการตัดคำที่นิยมใช้กันทั่วไปคือวิธีการเทียบคำที่ยาวที่สุด วิธีการเทียบคำที่สั้นที่สุด วิธีการตัดคำที่ใช้ความถี่ของคำและวิธีการย้อนรอยกลับ ซึ่งมีรายละเอียดดังต่อไปนี้

1. วิธีการเทียบคำที่ยาวที่สุด (Longest Word Pattern Matching) วิธีนี้จะทำการตรวจสอบสายอักขระ (String) ที่นำเข้ามาจากซ้ายไปขวา จากนั้นนำไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรม หากตรวจสอบพบว่าพบพยางค์มากกว่า 1 พยางค์ในพจนานุกรม จะทำการเลือกพยางค์ที่ยาวที่สุดแล้วทำต่อไปเรื่อยๆ จนจบสายอักขระ ตัวอย่างคำว่า “กองกลาง” การตัดคำโดยวิธีนี้จะนำสายอักขระไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรมจะพบคำว่า ก , กอ และคำว่า กอง ส่วนคำว่า กองก ไม่พบอยู่ในพจนานุกรม ดังนั้นจึงได้คำว่า กอง ซึ่งเป็นคำที่ยาวที่สุดที่หาพบ ส่วนที่เหลือคือ กลาง เมื่อนำไปค้นในพจนานุกรมจะได้ว่า ก , กล , กลาง ดังนั้นจึงเลือกคำว่า กลาง คำที่ได้จากการตัดคำโดยวิธีการนี้จึงเป็น กอง กลาง วิธีการนี้ให้ความถูกต้องหลังการตัดคำสูงกว่าวิธีการอื่นโดยเฉพาะเมื่อใช้ร่วมกับวิธีย้อนรอยกลับ

2. วิธีการเทียบคำที่สั้นที่สุด (Shortest Word Pattern Matching) วิธีการนี้คล้ายกับวิธีการเทียบคำที่ยาวที่สุด เพียงแต่จะเลือกคำที่สั้นที่สุดที่พบก่อน แต่วิธีนี้พบว่าได้จำนวนคำมากที่สุดแต่ความถูกต้องของคำหลังทำการตัดคำน้อยกว่าการใช้วิธีเทียบคำที่ยาวที่สุดตัวอย่างคำว่า “โคลงเรือ” การตัดคำโดยวิธีนี้จะเลือกเอาคำแรกที่ค้นหาเจอจากพจนานุกรม ดังนั้นจะได้ว่า โค ลง เรือ (โดยไม่เลือกคำว่า “โคลง” ที่จะพบต่อไปภายหลังหากทำการค้นหาต่อ) วิธีนี้ใช้เวลาน้อยกว่าการเทียบคำยาวที่สุด แต่ความถูกต้องที่ได้การตัดคำแบบเทียบคำยาวที่สุดจะมากกว่า

3. วิธีการตัดคำที่ใช้ความถี่ของคำ (Word Usage Frequency) วิธีการนี้เป็นแนวทางหนึ่งในการแก้ปัญหาคำกำกวมของประโยคภาษาไทย โดยการวิเคราะห์ความถี่ของการใช้คำในชีวิตประจำวัน โดยจัดเรียงคำในพจนานุกรมตามความถี่ที่พบ และใช้วิธีการตัดคำแบบการเทียบคำที่สั้นที่สุดมาตัดคำและนำไปตรวจสอบกับตารางที่สร้างขึ้นมา ตารางนี้จะมีคำและมีการจัดค่าความถี่ของคำแต่ละคำไว้ เช่นคำว่า ก็อด จะตัดไปเป็น ก็- อด หรือ ก็อด เนื่องจากการใช้งาน

ลักษณะคำในภาษาไทยจะมีการใช้คำว่า ก็-อด บ่อยกว่าคำว่า ก็อด ระบบจะแสดงผลลัพท์เป็นคำว่า ก็-อด

4. วิธีการย้อนรอยกลับ (Back Tracking) เมื่อทำการเปรียบเทียบคำที่นำมาตัดคำกับคำที่มีอยู่ในพจนานุกรม อาจพบกรณีที่คำที่พบมีมากกว่า 1 คำแล้วทำการเลือกคำที่ยาวที่สุดทำให้สายอักขระที่ตามมาจากคำนั้นไม่สามารถตัดคำได้เนื่องจากไม่พบตามพจนานุกรม กรณีนี้จะทำการย้อนไปอีกคำที่ไม่ถูกเลือกแล้วทำการตัดคำต่อไปตัวอย่างเช่นคำว่า “เมื่อยามนี้” การเปรียบเทียบกับพจนานุกรมจะได้ว่า เมื่อ, เมื่อย ดังนั้นจึงเลือกคำที่ยาวที่สุดจะได้คำว่าเมื่อย ส่วนที่เหลือคือ –ามนี้ ซึ่งไม่พบอยู่ในพจนานุกรม ดังนั้นจะทำการย้อนกลับไปเพื่อเลือกอีกคำหนึ่งคือ เมื่อ จะได้เป็น เมื่อ ยาม นี้ (โดยคำว่า ยามเกิดจากการเลือกคำที่ยาวที่สุดระหว่าง ยา และ ยาม)

2.5 ระบบฐานข้อมูลออนโทโลจีภาษาอังกฤษ WordNet

Varelas G., Voutsakis E., Raftopoulou P., Petrakis E. และ Milios E. (2005) ได้อธิบายไว้ว่า WordNet คือฐานข้อมูลที่รวบรวมคำศัพท์ภาษาอังกฤษไว้ แต่ไม่ได้ใช้งานในรูปแบบ Dictionary กล่าวคือไม่ได้สนใจแค่ว่าคำคำนี้แปลว่าอะไร แต่ WordNet จะเน้นไปที่ความสัมพันธ์ระหว่างคำศัพท์ โดยสามารถถือว่า WordNet เป็น Ontology อันหนึ่งที่รวบรวมคำศัพท์ไว้มากกว่า 100,000 คำ

WordNet ประกอบไปด้วยคำนาม (Nouns) คำกริยา (Verbs) คำคุณศัพท์ (Adjectives) และคำวิเศษณ์ (Adverbs) โดยคำที่มีความสัมพันธ์เกี่ยวข้องกันจะนำมาเกี่ยวโยงกันด้วย Synonym Sets (Synsets) โดยข้อมูลใน Synsets จะเกี่ยวโยงกันด้วย Senses คำหนึ่งคำที่มีความหมายมากกว่าหนึ่งความหมายจะมี Senses มากกว่าหนึ่ง Senses นอกจากนี้ WordNet ยังมีความสัมพันธ์ระหว่างคำนอกเหนือจาก Synonym อีก 2 รูปแบบ คือ ความสัมพันธ์แบบ Is-A หรือเรียกว่า Hyponym และ Hypernym และความสัมพันธ์แบบ Part-Of หรือเรียกว่า Meronym และ Holonym

ตัวอย่างความสัมพันธ์แบบ Is-A

สุนัข เป็น Hypernym ของ ดัลเมเชี่ยน

ดัลเมเชี่ยน เป็น Hyponym ของสุนัข

ข้อสังเกตหรือวิธีทำความเข้าใจแบบง่ายๆ Hypernym มีรากศัพท์คือ Hyper แปลว่าเหนือ เพราะฉะนั้น คำ ก. Hypernym ของคำ. ข แสดงว่าคำ ก. อยู่เหนือกว่าคำ ข. หรือคำ ก. มีความหมายกว้างๆ แต่คำ ข. มีความหมายแบบเฉพาะ

ตัวอย่างความสัมพันธ์แบบ Part-Of

ตึก เป็น Holonym ของหน้าต่าง

หน้าต่างเป็น Meronym ของตึก

ข้อสังเกตหรือวิธีทำความเข้าใจแบบง่ายๆ Meronym คล้ายกับคำว่า Member แปลว่าสมาชิก เพราะฉะนั้น คำ ก. เป็น Meronym ของ คำ ข. แสดงว่า คำ ก. เป็นสมาชิกหรือเป็นส่วนประกอบของ คำ ข.

คำที่มีความหมายคล้ายกัน หรือ Semantic Similarity จะสนใจในส่วนของ Synset และนอกจากนั้น ยังต้องเอา Hypernym และ Hyponym มาทำการวิจัยควบคู่ไปด้วย นอกจากนั้น การค้นหาคำที่มีความหมายคล้ายกันจะสนใจเฉพาะ Noun และ Verb เท่านั้น ซึ่งการค้นหาคำที่มีความหมายคล้ายกัน สามารถแบ่งได้เป็น 4 วิธีใหญ่ๆได้แก่

1. Edge Counting Methods วัดคำศัพท์ 2 คำ โดยใช้ความยาวของ Path ที่เชื่อมต่อแต่ละคำและตำแหน่งของคำในกลุ่ม
2. Information Content Methods วัดความแตกต่างของเนื้อหาของสองคำ โดยใช้ความเป็นไปได้ที่จะเกิดขึ้นในเอกสาร
3. Feature Based Methods วัดความคล้ายกันของคำสองคำ โดยดูที่ Properties ของคำ
4. Hybrid Methods รวมเอาวิธีทั้งหมดเข้าด้วยกัน

วิธีการหาความคล้ายของคำโดยส่วนมากจะค้นหาจาก Ontology ซึ่งสามารถแบ่งได้ 2 ประเภทคือ

1. Single Ontology คือ คำสองคำที่ต้องการหามาจาก Ontology เดียวกัน ใช้วิธี Edge Counting Methods และ Information Content Methods
2. Cross Ontology คือ คำสองคำที่ต้องการหามาจาก Ontology มากกว่าหนึ่ง Ontology ใช้วิธี Feature Based Methods และ Hybrid Methods

ตารางที่ 2.1 แสดงจำนวนคำศัพท์ในฐานความรู้เวิร์ดเน็ต

Part of Speech	Unique String	Synsets	Total Word Sense Pair
Noun (n)	114,648	79,689	146,273
Verb (v)	11,306	13,508	24,691
Adjective (a, s)	21,793	18,563	31,016
Adverb (r)	4,660	3,664	5,808
Total	152,407	115,424	207,788

จากตารางที่ 2.1 แสดงจำนวนข้อมูลทั้งหมดและประเภทคำต่างๆซึ่งแบ่งได้เป็น 3 กลุ่มดังนี้

1. Unique String คือกลุ่มของคำศัพท์ที่ถูกจัดหรือคัดออกมาให้เหลือเพียงหนึ่งคำจากคำศัพท์ที่ซ้ำกัน

2. Synsets คือกลุ่มของคำศัพท์ทั้งหมดที่มีควาหมายใกล้เคียงหรือเหมือนกันกับกลุ่มของ Unique String

3. Total Word Sense Pair คือกลุ่มคำศัพท์ทั้งหมดที่มีความหมายใกล้เคียงหรือเหมือนกันกับคำศัพท์ที่ไม่ได้ถูกจัดกลุ่มให้เหลือเพียงหนึ่งคำจากคำศัพท์ที่ซ้ำกัน

2.6 งานวิจัยที่เกี่ยวข้อง

วิศิษฎ์ วรรณภูมิ และศิพาณี นุชิตประสิทธิ์ชัย (2009) ได้ทำการศึกษาเรื่อง “การสืบค้นข้อมูลการให้บริการเว็บเซอร์วิสเชิงความหมาย” ได้นำเสนอแนวคิดเกี่ยวกับการนำแนวคิดเชิงความหมายและออนโทโลจีเข้ามาช่วยพัฒนาระบบสืบค้น โดยใช้การรวบรวมข้อมูลบริการของเว็บเซอร์วิสที่มีการมาลงทะเบียนไว้ แล้วนำมาสร้างเป็นฐานข้อมูลโดยอาศัยความสัมพันธ์ระหว่างข้อมูลต่างๆ ที่ได้ป้อนเข้าสู่ระบบ และใช้หลักการเชิงความหมายในการพัฒนาระบบการสืบค้นเมื่อถูกเรียกใช้งานจากผู้ใช้งาน โดยงานวิจัยนี้ได้มีการประเมินผลการทำงานโดยจากการทดลองใช้งานได้ผลลัพท์โดยรวมเป็นที่น่าพอใจของผู้วิจัย ซึ่งผลลัพท์ที่ได้จากการสืบค้นแสดงให้เห็นว่ามีความรวดเร็วและแม่นยำมากขึ้น

ธนกร หวังพิพัฒน์วงศ์, อานนท์ ไกรเสวกวิสัย และสราวุธิ ราษฎร์นิยม (2009) ได้ทำการศึกษาเรื่อง “ระบบค้นหารูปภาพโดยใช้หลักการเว็บเชิงความหมาย” และนำเสนอรูปแบบการค้นหารูปภาพสถานที่ท่องเที่ยวในประเทศไทย โดยนำหลักการเชิงความหมายมาช่วยในการพัฒนาระบบสืบค้นรูปภาพ ซึ่งผู้วิจัยได้รวบรวมคำบรรยายภาพของแต่ละภาพมาจัดลำดับความสัมพันธ์กัน

และได้นำฐานข้อมูลออนโทโลจี WordNet เข้ามาช่วยในการอ้างอิงความสัมพันธ์ของคำศัพท์ภาษาอังกฤษ ทำให้เกิดการค้นหาที่แม่นยำมากขึ้น จากการทดสอบประสิทธิภาพในการค้นหาได้พบว่าได้ค่าความแม่นยำ 6.6 จากคะแนนเต็ม 10 ซึ่งอยู่ในเกณฑ์ที่น่าพอใจ

Yi Jin, Zhuying Lin และ Hongwei Lin (2008) ได้ทำการศึกษาเรื่อง "The Research of Search Engine Based on Semantic Web" โดยได้นำเสนอการนำอัลกอริทึม TFIDF เข้ามาช่วยในการประมวลผลในการแยกคำหรือการตัดคำ ซึ่งเป็นการนำหลักการทางสถิติเข้ามาช่วย โดยการทำงานจะใช้การจัดทำดัชนีชี้วัดค่าความสำคัญของข้อมูล ที่ช่วยประเมินความสำคัญของข้อมูลในการนำมาแสดงผล เพื่อเพิ่มความถูกต้องในการสืบค้น ทำให้การสืบค้นมีการคัดกรองข้อมูลผลัพธ์ก่อนนำมาแสดงผลหลากหลายขึ้น จากการทดสอบระบบโดยทำการทดสอบเปรียบเทียบกับระบบสืบค้นที่ไม่ใช้อัลกอริทึม TFIDF ผลลัพธ์ที่ได้จากระบบที่พัฒนาขึ้นมีความถูกต้องแม่นยำมากกว่าระบบที่ไม่ได้ใช้อัลกอริทึม TFIDF

Alexnader Maedche และ Steffn Stabb (2001) ได้นำเสนอ "Ontology Learning for the Semantic Web" ซึ่งอธิบายถึงที่มาและหลักการออนโทโลจี โดยเน้นที่ลักษณะความสัมพันธ์ของลำดับชั้นข้อมูล แนวคิดของกฎการใช้งานออนโทโลจีและสถาปัตยกรรมของออนโทโลจีที่จะนำมาใช้กับระบบเว็บเชิงความหมาย ซึ่งในงานวิจัยนี้ผู้วิจัยได้เน้นในส่วนของพื้นฐานการพัฒนาออนโทโลจี และหลักการในการออกแบบความสัมพันธ์ระหว่างสิ่งต่างๆ เพื่อให้ผู้ที่สนใจจะเริ่มต้นพัฒนาออนโทโลจีสามารถนำไปเป็นหลักการพัฒนาออนโทโลจีได้โดยถูกต้อง

Tim Finin, et al. (2004) ได้นำเสนอ "Swoogle: a Semantic Web Search and Metadata Engine" ซึ่งเป็นระบบช่วยบริการสืบค้นข้อมูลออนโทโลจีผ่านเว็บไซต์ ที่พัฒนาโดยใช้หลักการเว็บเชิงความหมายซึ่งระบบการทำงานของระบบช่วยสืบค้นข้อมูลนี้จะมีลักษณะคล้ายๆ กับระบบช่วยสืบค้นข้อมูลทั่วไปแต่ผลลัพธ์ในการสืบค้นของเว็บไซต์นี้ จะเป็นผลลัพธ์ที่อยู่ในรูปของเอกสารที่เป็นเอกสารออนโทโลจี เช่น เอกสาร RDF RDFS OWL เป็นต้น โดยผู้พัฒนาเน้นนำเสนอตัวอย่างในการพัฒนาระบบโดยใช้หลักการเชิงความหมายในการค้นหาเอกสารออนโทโลจี โดยเอกสารออนโทโลจีที่สามารถค้นหานั้น ระบบได้นำมาจัดความสัมพันธ์ของแต่ละเอกสารไว้แล้วหรือก็คือการค้นเอกสารออนโทโลจีจากออนโทโลจีนั่นเอง เหมาะสำหรับผู้พัฒนาที่ต้องการหาตัวอย่างมาเพื่ออ้างอิงในการพัฒนาระบบต่อไป

Reza Hemayati, Weiyi Meng และ Clement Yu (2007) ได้นำเสนอ "Semantic-Based

Grouping of Search Engine Results Using WordNet" โดยได้นำหลักการจัดกลุ่มของคำมาช่วยเพิ่มประสิทธิภาพของผลลัพท์ในการสืบค้นข้อมูลสารสนเทศ ซึ่งในระบบนี้ได้นำเสนอ SRR GroupingAlgorithm ที่นำสองรูปแบบมาเปรียบเทียบคือ Largest Frequency of Use (LF) และ Largest Category (LC) ซึ่งระบบได้ทำการพัฒนาด้วยภาษาจาวาและใช้ JWNL (Jena WordNet Libraly) เป็นเครื่องมือในการเชื่อมต่อฐานข้อมูล WordNet ผลการทดสอบออกมาปรากฏว่าการใช้ SSR Algorithm ได้ค่าความถูกต้องแม่นยำที่สุดโดยได้ค่าเฉลี่ยสูงสุดอยู่ที่ 93% ส่วน LF และ LC ได้ค่าความถูกต้อง 75% และ 78% ตามลำดับ จากผลการทดลองสรุปว่าปัญหาในการจัดกลุ่มคำในการสืบค้นข้อมูลนั้น สามารถจัดการได้ด้วย SSR Algorithm ซึ่งจะได้ผลลัพท์ที่แม่นยำขึ้นกว่าการใช้หลักการทั่วไป

Junaidah Mohamed kassim และ Mahathir Rahmany (2009) ได้นำเสนอ "Introduction to Semantic Search Engine" โดยได้นำเสนอแนวคิดเกี่ยวกับหลักการของการพัฒนาระบบช่วยบริการสืบค้นโดยใช้หลักการเว็บเชิงความหมายมาช่วยพัฒนา ซึ่งในงานวิจัยนี้ได้อธิบายภาพรวมของหลักการและนำเสนอเทคโนโลยีที่จำเป็นต่อการพัฒนาระบบ และยังได้นำเสนอเว็บไซต์ที่ให้บริการสืบค้นสารสนเทศซึ่งใช้หลักการเชิงความหมาย สำหรับผู้ที่สนใจสามารถเข้าไปทดลองใช้บริการสืบค้น และสุดท้ายได้สรุปข้อแตกต่างระหว่างเทคโนโลยีเว็บเชิงความหมายซึ่งเป็นเทคโนโลยีเว็บ 3.0 กับเทคโนโลยีแบบปัจจุบันซึ่งเป็นเทคโนโลยี 2.0

ไพศาล เจริญพรสวัสดิ์ (1999) ได้นำเสนอ "Feature-based Thai word segmentation" โดยนำเสนอการตัดคำซึ่งเป็นคำที่อยู่นอกเหนือจากกฎเกณฑ์หลัก หรือการตัดคำที่ไม่ได้มีการยอมรับแต่มีการใช้กันนั่นเอง ซึ่งงานวิจัยนี้แบ่งปัญหาในการตัดคำออกเป็น 2 ชนิดด้วยกันคือ

1. ปัญหาความกำกวม

2. ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม

การแก้ปัญหาการตัดคำ ได้นำคุณลักษณะโดยการใช้การเรียนรู้ของเครื่อง 2 แบบ ได้แก่ ริปเปอร์และวินโนว์ ซึ่งนำคุณสมบัติและบริบทของคำที่อยู่รอบๆ มาใช้ในการแก้ปัญหาการตัดคำโดยใช้สถิติเข้ามาช่วยโดยจะสรุปออกมาเป็นสมการดังนี้

$$t_t \max \arg \text{PROVB}(C_1,...,C_t \mid w_1,...,w_1) = \tau$$

โดยที่ τ คือ C1 ,…Ct ที่ให้ค่าความน่าจะเป็นมีค่ามากที่สุด

Ci คือหน้าที่คำจองคำ Wi

Wi คือลำดับของประโยคหนึ่งๆ

เนื่องจากการตัดคำโดยวิธีนี้ต้องอาศัยหน้าที่ของคำ แต่เนื่องจากคำบางคำที่ไม่ปรากฏในพจนานุกรมหรือคำที่ยังไม่รู้จักจึงทำให้การตัดคำนี้ยังทำไม่ได้ทันที ต้องทำการกำหนดหน้าที่คำเสียก่อน

ดวงแก้ว สวามิภักดิ์ (1990) ได้นำเสนอ “การสร้างซอฟต์แวร์วิเคราะห์ไวยากรณ์ไทยภายใต้ระบบยูนิกซ์” ซึ่งงานวิจัยนี้ได้สร้างซอฟต์แวร์วิเคราะห์ไวยากรณ์ไทยภายใต้ระบบยูนิกซ์ เป็นงานวิจัยการตัดคำภาษาไทยโดยใช้กฎที่สร้างขึ้นเองจำนวน 43 กฎ และมีการนำพจนานุกรมเข้ามาร่วมด้วย โดยสาเหตุที่นำทั้งกฎไวยากรณ์และพจนานุกรมเข้ามาช่วยในการตัดคำนั้นก็เพื่อจะแก้ไขปัญหาการตัดคำโดยใช้พจนานุกรมเพียงอย่างเดียว ซึ่งไม่สามารถตัดคำได้อย่างถูกต้องในกรณีที่คำนั้นไม่มีอยู่ในพจนานุกรม งานวิจัยนี้ได้ทำภายใต้ระบบปฏิบัติการยูนิกซ์และได้มีการนำโปรแกรมเล็กซ์ (Lex) เข้ามาร่วมด้วย โดยกฎที่ได้มานี้จะไม่มีการรวมตัวสะกดเข้าไปในกฎด้วย ยกเว้นบางกรณี เนื่องจากโปรแกรมเล็กซ์จะพยายามสร้างกลุ่มอักษร (Token) ที่ยาวที่สุดก่อน ดังนั้นหากมีการนำตัวสะกดเข้ามาใช้ จะเป็นสาเหตุให้มีการรวมเอาอักษรตัวหน้าของคำถัดไปมาเป็นตัวสะกดได้ ซึ่งเมื่อได้มีการวิเคราะห์ด้วยกฎแล้ว ขั้นตอนต่อไปจะมีการรวมกลุ่มตัวอักษรเข้าด้วยกัน โดยทำการตรวจสอบจากพจนานุกรม ส่วนโครงสร้างของพจนานุกรมที่นำมาใช้เป็นฐานข้อมูลแบบรีเลชัน (Relation DBMS) ซึ่งใช้คำเป็นดรรชนี (Index) และไฟล์ดรรชนีได้พัฒนาขึ้นโดยใช้โครงสร้างข้อมูลแบบบี-ทรี (B-Tree)

2.7 บทสรุป

จากการศึกษาปริทัศน์ วรรณกรรมและงานวิจัยที่เกี่ยวข้องดังที่กล่าวมาข้างต้นนั้น จะเห็นได้ว่าปัญหาส่วนใหญ่ที่เกิดขึ้นจากการใช้งานระบบช่วยบริการสืบค้นข้อมูลนั้น เกิดจากการที่ระบบส่วนใหญ่ยังมีปัญหาในการสืบค้นข้อมูลภาษาไทย ซึ่งก่อให้เกิดปัญหาในการแสดงผลลัพธ์ได้ไม่ตรงตามที่ผู้ใช้งานต้องการสืบค้น เนื่องจากคำในภาษาไทยบางคำสามารถมีได้หลายความหมาย ในงานวิจัยนี้ได้นำหลักการเชิงความหมายและนำหลักการตัดคำในภาษาไทย มาช่วยพัฒนาในระบบช่วยบริการสืบค้นสารสนเทศ ซึ่งสามารถช่วยอธิบายความหมายของคำหลักในการสืบค้นได้ดียิ่งขึ้น อีกทั้งยังนำการแก้ไขปัญหาของการสืบค้นด้วยคำหลักที่ออกเสียงมาจากภาษาอังกฤษ ซึ่งมักมีการเขียนที่ไม่ถูกต้อง นอกจากนี้ข้อแตกต่างของงานวิจัยนี้กับงานวิจัยอื่นคือ งานวิจัยนี้ได้มุ่งเน้นพัฒนาระบบบริการสืบค้นสารสนเทศโดยพยายามที่จะพัฒนาการค้นหาข้อมูลในภาษาไทยเป็นสำคัญ โดยจะเน้นการนำเสนอการให้บริการสืบค้นสารสนเทศซึ่งเป็นข้อมูลในด้านเทคโนโลยีสารสนเทศ ทำ

ให้การสืบค้นข้อมูลในภาษาไทยทำได้สะดวกมากยิ่งขึ้น เพื่อให้การพัฒนาระบบให้บริการสืบค้นสารสนเทศมีความสมบูรณ์ยิ่งขึ้นในส่วนของการจัดความสัมพันธ์ของคำหลักในการค้นหา ซึ่งในส่วนนี้ได้อาศัยภาษาหลักในการพัฒนาคือ ภาษาจาวา ในการสร้างฐานความรู้ออนโทโลจีจะใช้โปรแกรมโปรเทเจ ซึ่งมีส่วนการติดต่อผู้ใช้งานเป็นแบบกราฟิกและมีเครื่องมือสำหรับสร้างโดเมนของออนโทโลจี และรูปแบบข้อมูลที่สะดวกในการป้อนข้อมูลรองรับการขยายระบบงาน สามารถทำงานบนเครือข่ายได้

ในบทต่อไปผู้วิจัยจะนำเสนอระเบียบวิธีวิจัยและกรอบแนวคิด โดยจะแสดงถึงขั้นตอนในการวิจัยและพัฒนาระบบ โดยจะเริ่มจากระเบียบวิธีการวิจัย แนวทางการศึกษาข้อมูล โครงสร้างโดยรวมของระบบค้นคืนสารสนเทศ ผังการทำงานโดยรวมของระบบ การวิเคราะห์ความต้องการของระบบ การออกแบบระบบ แนวทางในการพัฒนาระบบและแนวทางและวิธีการทดสอบระบบ

บทที่ 3
ระเบียบวิธีวิจัยและกรอบแนวคิด

ในงานวิจัยนี้ ผู้วิจัยได้นำเสนอแนวทางในการแก้ปัญหาการพัฒนาระบบค้นคืนสารสนเทศและการตัดคำภาษาไทยสำหรับการค้นคืนสารสนเทศ โดยได้นำหลักการเชิงความหมายเข้ามาช่วยแก้ปัญหา เพื่อแก้ปัญหาในด้านประสิทธิภาพในส่วนผลลัพธ์ของระบบค้นคืนสารสนเทศด้านเทคโนโลยีสารสนเทศภาษาไทย เนื่องจากในปัจจุบันการค้นคืนสารสนเทศด้านเทคโนโลยีสารสนเทศภาษาไทยยังเกิดปัญหาในด้านความถูกต้องแม่นยำของผลลัพธ์อยู่เป็นอย่างมาก ในการแก้ปัญหานี้ผู้วิจัยมีระเบียบวิธีการวิจัยโดยเริ่มจาก ระเบียบวิธีการวิจัย แนวทางการศึกษาข้อมูล โครงสร้างโดยรวมของระบบค้นคืนข้อมูล ผังการทำงานโดยรวมของระบบ การวิเคราะห์ความต้องการของระบบ การออกแบบระบบ แนวทางในการพัฒนาระบบและแนวทางและวิธีการทดสอบระบบ

3.1 ระเบียบวิธีการวิจัย

ในการค้นคว้าวิจัยนี้ผู้วิจัยจะแบ่งวิธีการวิจัยออกเป็นขั้นตอนดังนี้

3.1.1 ศึกษาและรวมรวมงานวิจัยที่เกี่ยวข้อง

3.1.2 ศึกษารายละเอียดเกี่ยวกับวิธีการตัดคำภาษาไทยในรูปแบบต่างๆ และศึกษารายละเอียดระบบค้นคืนข้อมูล เพื่อนำมาเป็นโครงสร้างหลักของระบบ

3.1.3 ศึกษาและรวบรวมข้อมูลเกี่ยวกับหลักการเชิงความหมาย

3.1.4 ศึกษาและรวบรวมข้อมูลในการพัฒนาเว็บแอพพลิเคชันเพื่อนำมาใช้กับหลักการเชิงความหมาย

3.1.5 ออกแบบโครงสร้างโดยรวมของระบบ

3.1.6 ทำการพัฒนาระบบตามที่ได้ออกแบบไว้

3.1.7 ทดลองแก้ไขปรับปรุงระบบ

3.1.8 วิเคราะห์และสรุปผลการวิจัย

3.2 แนวทางการศึกษาข้อมูล

การศึกษาข้อมูล ความรู้ แนวคิด หลักการ วิธีการปฏิบัติและข้อเสนอแนะต่างๆ เพื่อนำมาประกอบการดำเนินการวิจัยในครั้งนี้ ผู้วิจัยได้ศึกษาข้อมูลจากแหล่งต่างๆ เช่น ตำราเรียน วารสาร หนังสือ วิทยานิพนธ์ บทความวิจัยและเว็บไซต์ทั้งเว็บไซต์ของประเทศไทยและต่างประเทศในด้านที่เกี่ยวข้องกับระบบค้นคืนสารสนเทศ ด้านการตัดคำและด้านการพัฒนาเทคโนโลยีเชิงความหมาย จากนั้นได้นำองค์ความรู้มาต่างๆ ที่ได้รวบรวมมาพัฒนาระบบ

ในการศึกษาข้อมูลผู้วิจัยได้ทดลองนำ ส่วนเสริมของภาษา Java ที่ชื่อว่า Lucene 3.0 (http://www.lucene.apache.org) มาศึกษาและใช้งาน เนื่องจากพบว่ามีความเหมาะสมต่อการใช้งานในส่วนของการพัฒนาระบบค้นคืนสารสนเทศ โดยมีหน่วยงานใหญ่ๆ นำไปใช้งานในการพัฒนาส่วนค้นคืนสารสนเทศอย่างแพร่หลาย เช่น Apple inc. IBM เป็นต้น จากการทดลองใช้งานทำให้ผู้วิจัยสามารถมั่นใจได้ว่าส่วนเสริมนี้มีความสามารถครอบคลุมสามารถนำมาใช้ในงานวิจัยชิ้นนี้ได้เป็นอย่างดี อีกทั้ง Lucene 3.0 เป็นซอฟต์แวร์เสรีที่พัฒนาโดย Apache Foundation ซึ่งถูกพัฒนาให้สามารถนำไปพัฒนาระบบค้นคืนสารสนเทศได้ในหลากหลายภาษา เช่น Java PHP เป็นต้น ทำให้สามารถนำมาใช้พัฒนาระบบในงานวิจัยได้โดยสะดวกไม่ติดปัญหาด้านลิขสิทธ์ และจากการศึกษาข้อมูลผู้วิจัยสามารถสรุปข้อดีหลักๆ ของ Lucene3.0 ในการทดลองใช้งานได้ดังนี้

- ใช้ทรัพยากรหน่วยความจำน้อยเมื่อเทียบกับเทคโนโลยีที่ช่วยในการพัฒนาระบบค้นคืนสารสนเทศอื่นๆ โดยต้องการหน่วยความจำในการทำงานเริ่มต้นเพียงแค่ 1 Megabytes เท่านั้น
- สามารถสร้างดัชนีได้สะดวกและรวดเร็วอีกทั้งยังใช้เนื้อที่ในการเก็บดัชนีเพียงแค่ 20% – 30% เมื่อเทียบกับขนาดของเอกสารที่นำมาสร้างดัชนี
- สามารถค้นคืนสารสนเทศได้จากหลายส่วน เช่น ชื่อสารสนเทศ เนื้อหา ผู้แต่ง เป็นต้น โดยข้อมูลในแต่ละส่วนที่ค้นคืนนั้นจะค้นหาจากดัชนีที่ทำการสร้างไว้ ซึ่งหมายถึงผู้พัฒนาต้องทำการจัดทำดัชนีสำหรับการค้นคืนสารสนเทศในแต่ละประเภทตามที่ผู้ใช้งานต้องการ
- มีคลาสที่ทำหน้าที่จัดการเกี่ยวกับการแสดงผลลัพธ์ให้เลือกใช้งานมากมาย ทำให้ง่ายต่อการจัดการและการแก้ไขโปรแกรม อีกทั้งยังมีส่วนเสริมให้สามารถนำมาใช้งานร่วมกับ Lucene3.0 ได้ตามแต่ละวัตถุประสงค์ของการใช้งาน
- ทำงานร่วมกับภาษา Java ได้เป็นอย่างดีทำให้ไม่เกิดปัญหาเมื่อมีการนำไปใช้งานข้ามระบบปฏิบัติการ

ส่วนในการศึกษาข้อมูลเกี่ยวกับการตัดคำภาษาไทยนั้น ข้อมูลในการพัฒนาระบบตัดคำในภาษาไทยส่วนใหญ่นั้น มักจะเป็นการพัฒนาเพื่อการนำไปใช้งานเฉพาะด้าน คือเน้นการพัฒนาเพื่อใช้ในงานวิจัยของนักวิจัยแต่ละท่าน เมื่อนำมาใช้กับงานวิจัยอื่นๆ มักจะใช้งานได้ไม่ดีนัก เช่น ด้านพืชศาสตร์ หรือด้านการท่องเที่ยว เป็นต้น ทำให้ผู้วิจัยได้เกิดแนวคิดในส่วนของการพัฒนาเพื่อนำมาใช้ในระบบค้นคืนสารสนเทศ โดยผู้วิจัยได้เลือกข้อดีข้อเสียของหลักการตัดคำแต่ละรูปแบบมาคัดเลือกและทดลองเพื่อนำมาใช้งาน ในการเลือกวิธีการในการตัดคำนั้นผู้วิจัยได้ทำการศึกษาเอกสารที่เกี่ยวข้องกับด้านเทคโนโลยีสารสนเทศภาษาไทย ว่ามีการใช้คำส่วนใหญ่ในรูปแบบใดเพื่อให้สามารถเลือกใช้วิธีการได้ถูกต้องและให้ผลลัพธ์ได้ดีที่สุด เมื่อคัดเลือกวิธีการได้แล้วผู้วิจัยจะนำมาพัฒนาต่อเพื่อให้เหมาะสมกับงานวิจัยอีกครั้งหนึ่ง

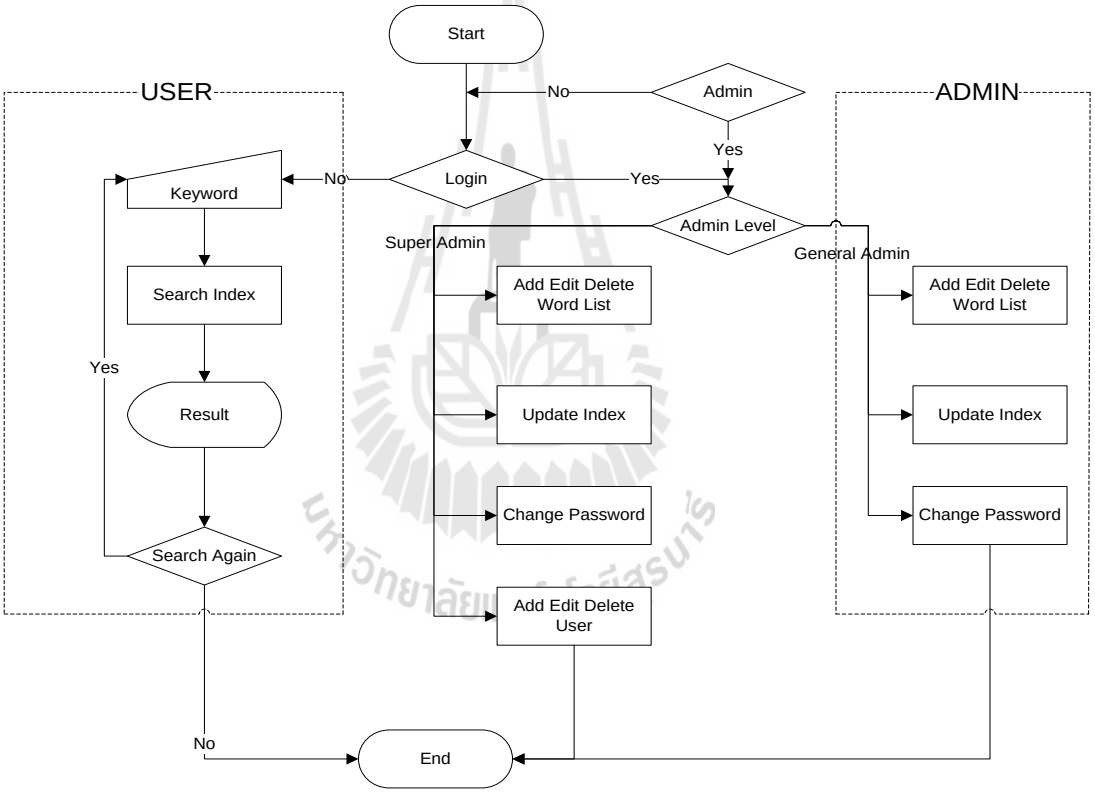
ในด้านของการพัฒนาการค้นคืนสารสนเทศ เพื่อให้เหมาะสมกับการใช้งานในภาษาไทยนั้น ผู้วิจัยเห็นว่าเนื่องจากความหมายของคำในภาษาไทยบางคำอาจมีได้หลายความหมาย ในการค้นคืนสารสนเทศอาจเกิดการแสดงผลลัพธ์ได้ไม่ตรงตามความหมายที่ผู้ใช้งานต้องการ ในส่วนนี้ผู้วิจัยได้ศึกษาข้อมูล และเห็นว่าการนำเทคโนโลยีเชิงความหมายเข้ามาพัฒนาร่วมกับระบบค้นคืนสารสนเทศ จะทำให้ผลลัพธ์ในการค้นคืนมีความแม่นยำมากขึ้น เนื่องจากเทคโนโลยีเชิงความหมายสามารถอธิบายความหมายของคำแต่ละคำ ในรูปแบบความสัมพันธ์ไม่ได้เน้นไปที่การเปรียบเทียบคำที่เหมือนกัน เพราะภาษาไทยบางคำสามารถมีได้หลายความหมายซึ่งขึ้นอยู่กับรูปแบบการเขียนและรูปแบบในการนำมาเรียงเป็นประโยค เมื่อนำหลักการเชิงความหมายมาใช้งานร่วมกับการค้นคืนสารสนเทศแล้วจะทำให้ระบบสามารถเข้าใจความหมายของคำหลักในการค้นคืนข้อมูลได้ดีขึ้นและสามารถแสดงผลลัพธ์ได้ตรงตามความหมายที่ผู้ใช้งานต้องการมากขึ้น

หลังจากที่ผู้วิจัยได้ศึกษาและค้นคว้าข้อมูลความรู้จากแหล่งต่างๆ และได้ศึกษาความเหมาะสมทางด้านเทคนิคว่าซอฟต์แวร์ที่ใช้พัฒนา มีความเหมาะสมทางด้านระบบปฏิบัติการสามารถช่วยพัฒนาในการใช้งานได้ดีขึ้นเพียงใด ศึกษาด้านเวลาว่าใช้เวลาในการพัฒนาระบบมากน้อยเพียงใดและเมื่อนำมาใช้งานจะสามารถลดเวลาในการทำงานลงได้หรือไม่ ประสิทธิภาพของระบบเหมาะที่จะนำมาใช้กับข้อมูลจำนวนมากๆ ได้อย่างดีหรือไม่ที่สำคัญคือการใช้งานในภาษาไทยจะทำให้ได้ผลลัพธ์ของการค้นคืนสารสนเทศภาษาไทยได้อย่างมีประสิทธิภาพดีขึ้นจริง

3.3 โครงสร้างโดยรวมของระบบค้นคืนสารสนเทศ

เนื้อหาในส่วนนี้กล่าวถึงโครงสร้างโดยรวมของระบบค้นคืนข้อมูล โดยจะแสดงให้เห็นถึงภาพรวมในการทำงานของระบบค้นคืนข้อมูลที่เกิดขึ้นในการค้นคืนข้อมูลครั้งหนึ่งๆ ดังแสดงในรูปที่ 3.1

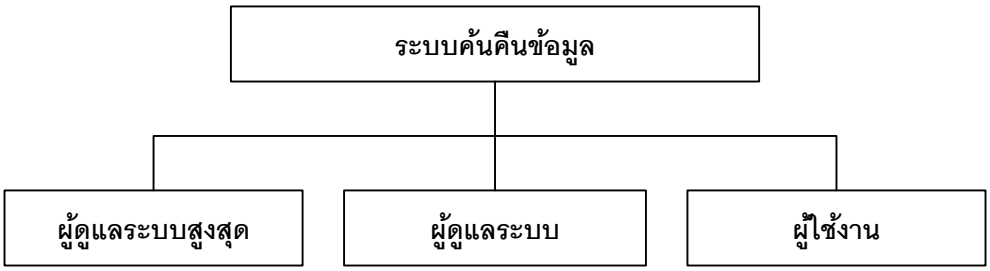
3.3.1 แผนภาพโดยรวมของระบบค้นคืนข้อมูล



รูปที่ 3.1 แสดงแผนภาพโดยรวมของระบบค้นคืนข้อมูล

3.4 ผังการทำงานโดยรวมของระบบ

ผังการทำงานนี้จะแสดงให้เห็นถึงผู้ที่มีส่วนเกี่ยวข้องกับระบบค้นคืนข้อมูล ว่าภายในระบบประกอบไปด้วยผู้ใดบ้างและจะแบ่งแยกหน้าที่ของแต่ละส่วนไปตามลำดับ ดังแสดงในรูปที่ 3.2



รูปที่ 3.2 แสดงผังการทำงานโดยรวมของระบบค้นคืนข้อมูล

3.4.1 ผังการทำงานของผู้ดูแลระบบสูงสุด Super Administrator

ผังการทำงานนี้จะแสดงให้เห็นถึงกิจกรรมต่างๆ โดยรวมที่เกิดขึ้นของผู้ดูแลระบบสูงสุด ดังแสดงในรูปที่ 3.3

รูปที่ 3.3 แสดงผังการทำงานที่เกิดขึ้นของผู้ดูแลระบบสูงสุด

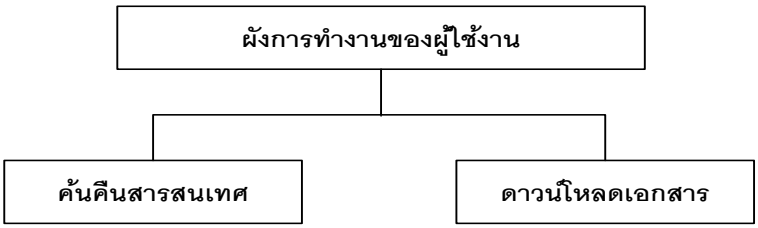
3.4.2 ผังการทำงานของผู้ดูแลระบบ General Administrator

ผังการทำงานนี้จะแสดงให้เห็นถึงกิจกรรมต่างๆ โดยรวมที่เกิดขึ้นของผู้ดูแลระบบทั่วไป ดังแสดงในรูปที่ 3.4

รูปที่ 3.4 แสดงผังการทำงานที่เกิดขึ้นของผู้ดูแลระบบ

3.4.3 ผังการทำงานของผู้ใช้งาน User

ผังการทำงานนี้จะแสดงให้เห็นถึงกิจกรรมต่างๆ ที่เกิดขึ้นของผู้ใช้งานค้นคืนสารสนเทศ ดังแสดงในรูปที่ 3.5



รูปที่ 3.5 แสดงผังการทำงานที่เกิดขึ้นของผู้ใช้งาน

3.5 การวิเคราะห์ความต้องการของระบบ

การวิเคราะห์ความต้องการของระบบ ผู้วิจัยเน้นวิเคราะห์จากปัญหาการพัฒนาและการใช้งานเว็บไซต์สำหรับค้นคืนสารสนเทศซึ่งเกิดขึ้นเป็นประจำในปัจจุบัน เช่นความยุ่งยากในการใช้งาน แสดงผลลัพธ์ได้ไม่ครอบคลุมจากการค้นคืนสารสนเทศและการปรับปรุงฐานข้อมูลในการค้นคืนสารสนเทศแต่ละครั้งใช้เวลานาน อีกทั้งยังใช้เนื้อที่ในการเก็บข้อมูลของระบบเป็นจำนวนมาก ผู้วิจัยพิจารณาว่าวิธีแก้ปัญหาคือต้องพัฒนาระบบที่มีหน้าจอแสดงผลที่เข้าใจง่าย ใช้งานได้สะดวกทั้งผู้ใช้งานและผู้ดูแลระบบ สามารถทำการปรับปรุงระบบฐานข้อมูลในการค้นคืนเอกสารได้รวดเร็วถูกต้อง มีการแสดงผลลัพธ์ที่แม่นยำและใช้ทรัพยากรในการประมวลผลของการค้นคืนสารสนเทศน้อย ซึ่งสามารถวิเคราะห์ได้เป็นหัวข้อหลักๆ ดังนี้

1. การวิเคราะห์ความต้องการของทรัพยากรระบบ

- ในด้านประสิทธิภาพของทรัพยากรเครื่องเซิฟร์เวอร์นั้น งานวิจัยนี้ผู้วิจัยได้นำส่วนเสริมของภาษา Java ที่ชื่อว่า Lucene เข้ามาใช้ในการพัฒนาระบบค้นคืนสารสนเทศ ซึ่งข้อดีของ Lucene นั้นนอกจากจะสามารถพัฒนาระบบค้นคืนสารสนเทศได้สะดวกรวดเร็ว แล้วยังต้องการทรัพยากรในการทำงานน้อยมากเมื่อเทียบกับส่วนเสริมอื่นๆ โดยมีความต้องการทรัพยากรหน่วยความจำในการทำงานเพียง 1 MB ซึ่งในการนำระบบไปติดตั้งใช้งานบนเซิฟร์เวอร์ไม่จำเป็นต้องใช้เครื่องเซิฟร์เวอร์ที่มีประสิทธิภาพสูงและราคาสูงนัก

- ในด้านของหน่วยความจำหลักนั้น ทางส่วนเสริม Lucene ได้มีการพัฒนาประสิทธิภาพในการจัดสร้างดัชนีสำหรับการค้นหาจากเอกสารต้นฉบับให้มีการใช้เนื้อที่ในการจัดเก็บน้อยลงถึง 70 เปอร์เซ็นต์เมื่อเทียบกับขนาดของเอกสารต้นฉบับ ทำให้ประหยัดเนื้อที่ของหน่วยความจำหลักได้เป็นอย่างมา สามารถลดต้นทุนในการจัดหาเนื้อที่ในการจัดเก็บสำหรับเครื่องเซิฟร์เวอร์ได้ดี

- ในด้านระบบปฏิบัติการนั้น Lucene มีพื้นฐานในการพัฒนามาจากภาษา Java ทำให้ไม่มีปัญหาในการทำงานข้ามระบบปฏิบัติการ โดยสามารถติดตั้งบนเซิฟร์เวอร์ที่ทำงานได้ในทุกระบบปฏิบัติการ

-ในส่วนของฐานข้อมูลหลักในการจัดเก็บคำสำคัญในการค้นคืนสารสนเทศ และรายชื่อผู้ใช้งานต่างๆ ผู้วิจัยได้เลือกใช้การจัดเก็บในรูปแบบเอกสาร XML เพื่อลดปริมาณการทำงานของระบบลงทำให้ระบบสามารถทำงานได้รวดเร็วขึ้น และใช้เนื้อที่ในการจัดเก็บฐานข้อมูลเพียงเล็กน้อย

2. การวิเคราะห์ความต้องการในการใช้งานของผู้ดูแลระบบ

ในการวิเคราะห์ความต้องการระบบการใช้งานของผู้ดูแลระบบนั้น ผู้วิจัยเน้นความเรียบง่ายและสะดวกในการใช้งาน ในส่วนของการเข้าสู่ระบบและปรับปรุงรายชื่อผู้ดูแลระบบ ผู้วิจัยจะจัดให้มีหน้าต่างเพื่อใส่รหัสผ่าน เพื่อให้ผู้ดูแลระบบสามารถเข้าสู่ระบบเพื่อปรับปรุงรายชื่อผู้ดูแลระบบ โดยผู้วิจัยจะจัดวางส่วนของการเข้าระบบไว้ในตำแหน่งที่ชัดเจนสวยงาม ผู้ดูแลระบบสามารถสังเกตเห็นได้ง่าย อีกทั้งในส่วนของการปรับปรุงรายชื่อผู้ดูแลระบบนั้น ผู้วิจัยได้ออกแบบการแสดงผลไว้ในรูปแบบตารางทำให้สังเกตและค้นหารายชื่อได้ง่ายและรวดเร็ว

3. การวิเคราะห์ความต้องการในการใช้งานของผู้ใช้งาน

ในการวิเคราะห์ความต้องการระบบของผู้ใช้งานนั้น ผู้วิจัยได้เน้นที่การใช้งานง่ายและสะดวกในการค้นคืนสารสนเทศ โดยในส่วนของผู้ใช้งานนั้นจะมีส่วนหลักที่สำคัญคือช่องใส่คำหลักในการค้นคืนสารสนเทศ ซึ่งผู้วิจัยจะออกแบบให้มีการใส่ส่วนช่วยสะกดคำในการพิมพ์เพื่อให้ง่ายต่อการพิมพ์คำที่ออกเสียงเลียนเสียงคำต่างประเทศ อีกทั้งในหน้าแสดงผลยังเน้นออกแบบให้มีการแสดงผลในรูปแบบตารางสลับสี เพื่อให้ผู้ใช้งานสามารถเลือกดูเอกสารที่แสดงจากผลลัพธ์ได้อย่างรวดเร็วอีกทั้งยังจะออกแบบให้มีการดาวน์โหลดเพื่อเก็บไว้ศึกษาได้อีกด้วย

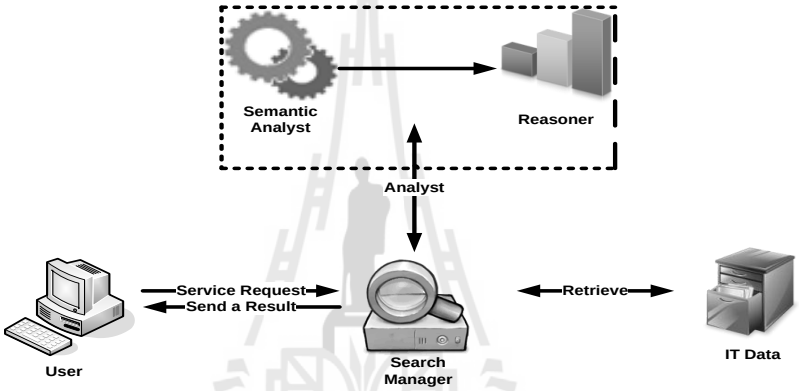
ซึ่งโดยรวมแล้ว ในส่วนของการวิเคราะห์ความต้องการของระบบที่ผู้วิจัยได้วิเคราะห์และสรุปออกมานั้น ทางด้านความต้องการของระบบนั้นมีความต้องการทรัพยากรของเครื่องเซิฟร์เวอร์น้อยมาก ซึ่งไม่มีปัญหาในการจัดหาเครื่องเซิฟร์เวอร์มาใช้งานเนื่องจากสามารถใช้งานกับเครื่องคอมพิวเตอร์ในปัจจุบันได้เป็นส่วนใหญ่ ในส่วนของการใช้งานนั้น ผู้วิจัยได้วิเคราะห์ว่าปัญหาในการใช้งานระบบค้นคืนสารสนเทศส่วนใหญ่นั้น เกิดจากการทำงานที่ช้าของระบบการใช้งานที่ยุ่งยากและปัญหาในการแสดงผลลัพธ์ที่ขาดความแม่นยำ จึงได้แนวคิดในการออกแบบระบบให้มีการทำงานได้รวดเร็วขึ้นใช้งานง่ายอีกทั้งยังแสดงผลลัพธ์ได้แม่นยำมากขึ้นกว่าเดิม

3.6 การออกแบบระบบ

เมื่อผู้วิจัยได้วิเคราะห์ความต้องการของระบบแล้ว ผู้วิจัยได้ทำการออกแบบระบบในส่วนต่างๆ ซึ่งมีวัตถุประสงค์เพื่อให้การพัฒนาระบบหรือการเขียนโปรแกรมทำได้ง่าย เป็นลำดับขั้นตอนและสามารถเห็นภาพรวมทั้งหมดของระบบได้อย่างชัดเจน ในการออกแบบระบบผู้วิจัยได้จัดทำแผนภาพการทำงานของระบบโดยคร่าวๆ เพื่อให้สามารถมองภาพรวมได้และอาจจะมีการแก้ไขในส่วนของรายละเอียดระบบอีกครั้งเมื่อได้เข้าสู่ระยะการพัฒนาระบบ ซึ่งอาจมีการเพิ่มเติมหรือลดในส่วนต่างๆ แต่ในบทนี้จะเป็นการแสดงแผนภาพระบบที่สำคัญของการพัฒนาระบบนี้ ส่วนการออกแบบวิธีการทดสอบระบบนั้นมีวัตถุประสงค์เพื่อให้ทดสอบระบบในส่วนต่างๆ อย่าง

ครบถ้วน และให้โปรแกรมใช้งานได้ตามความต้องการและมีประสิทธิภาพสูงสุด โดยได้นำผลที่ได้จากการวิเคราะห์ระบบในข้อ 3.2 มาเป็นหลักในการออกแบบ

การออกแบบระบบนั้น จะออกแบบให้มีการทำงานในรูปแบบของเว็บแอพพลิเคชัน เนื่องจากเหมาะสมกับการใช้งานและการพัฒนา อีกทั้งยังทำให้การค้นคืนในระบบเครือข่ายคอมพิวเตอร์ (Local Area Network) เป็นไปได้อย่างดีและใช้งานง่าย โดยในการออกแบบระบบผู้วิจัยได้ออกแบบระบบการทำงานโดยรวมดังแสดงในรูปที่ 3.6



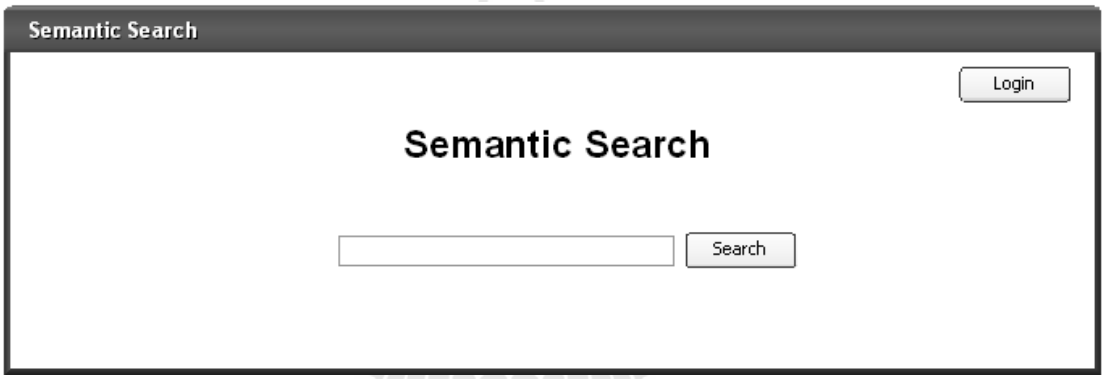
รูปที่ 3.6 แสดงการทำงานโดยรวมของระบบ

จากรูปที่ 3.6 แสดงให้เห็นถึงการทำงานโดยรวมของระบบ ซึ่งพัฒนาระบบบนหลักการเว็บแอพพลิเคชัน ในการทำงานจะมีส่วนสำคัญคือส่วนของการวิเคราะห์เชิงความหมาย ซึ่งจะทำงานร่วมกับระบบค้นคืนเพื่อคัดเลือกผลลัพธ์ที่ถูกต้องที่สุดสำหรับการค้นคืน พื้นฐานส่วนใหญ่ของระบบที่ผู้วิจัยได้ออกแบบระบบคือให้มีหลักการทำงานแบบ Client – Server และจะมีส่วนที่เป็นออนโทโลจีที่พัฒนาขึ้นมาเป็นตัวคัดกรองข้อมูลที่ถูกต้อง โดยเมื่อมีการเรียกใช้งานจากฝั่ง Client ฝั่ง Server จะมีการทำงานโดยการตัดคำหลักจากคำหลักที่ผู้ใช้ส่งมา แล้วทำการเปรียบเทียบกับออนโทโลจีที่อยู่ในระบบว่าคำหลักที่ผ่านการตัดมาแล้วมีความหมายคล้ายคลึงกับคำใดบ้างหลังจากนั้นระบบจะทำการคัดกรองข้อมูลที่ถูกต้องที่สุดแล้วทำการส่งผลลัพธ์กลับคืนให้ผู้ใช้งาน

3.6.1 ตัวอย่างหน้าจอของระบบ

ในการออกแบบหน้าจอของระบบ ผู้วิจัยได้เน้นที่ความเรียบง่ายในการใช้งาน จึงได้มีการออกแบบหน้าจอของระบบโดยแสดงแค่ส่วนที่จำเป็นในการใช้งานเท่านั้น โดยหน้าจอที่ออกแบบไว้จะมีด้วยกันสามรูปแบบ คือ รูปแบบหน้าจอของผู้ใช้งาน (User) หน้าจอของผู้ดูแลระบบและหน้าจอของผู้ดูแลระบบสูงสุด

1. รูปแบบหน้าจอของผู้ใช้งาน (User) ในการออกแบบหน้าจอใช้งานของผู้ใช้งานทั่วไป ผู้วิจัยได้คำนึงถึงความเรียบง่ายและความสะดวกในการใช้งาน โดยจากที่ผู้วิจัยได้กำหนดสิทธิสำหรับผู้ใช้งานทั่วไปไว้ที่การค้นคืนสารสนเทศเท่านั้น ทำให้หน้าจอของผู้ใช้งานทั่วไปมีลักษณะดังแสดงในรูปที่ 3.7



รูปที่ 3.7 แสดงรูปแบบหน้าจอของผู้ใช้งาน

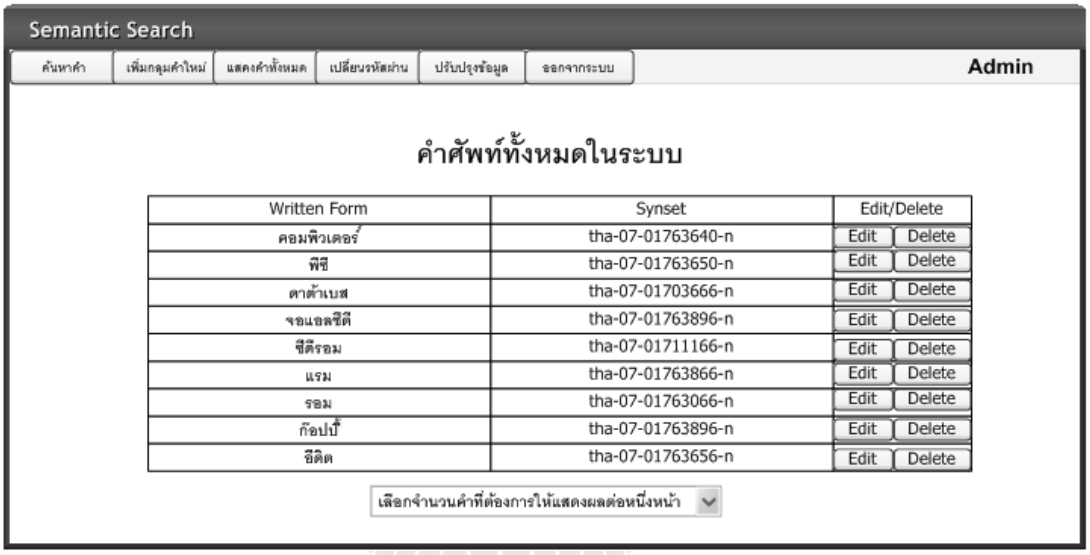
จากรูปที่ 3.7 จะเห็นว่าในหน้าเริ่มต้นใช้งานนั้น จะแสดงช่องใส่คำหลักในการค้นหาปุ่มค้นหาและปุ่มเข้าสู่ระบบเท่านั้น เมื่อผู้ใช้งานต้องการค้นคืนข้อมูลที่เกี่ยวข้องกับเรื่องใด ผู้ใช้งานจะต้องพิมพ์คำหลักในการค้นหาลงในช่องใส่คำหลัก แล้วคลิกที่ปุ่ม Search เพื่อเรียกให้ระบบทำงาน โดยระบบจะทำงานโดยการนำคำหลักที่ผู้ใช้ใส่เข้ามาเพื่อคัดกรองและแสดงผลลัพธ์ต่อไป โดยหน้าแสดงผลลัพธ์จะแสดงดังรูปที่ 3.8 สำหรับปุ่มเข้าสู่ระบบ (Login) จะใช้งานได้เฉพาะผู้ดูแลระบบสูงสุดและผู้ดูแลระบบเท่านั้น เนื่องจากต้องมีการระบุชื่อผู้ใช้งานและรหัสผ่าน

รูปที่ 3.8 แสดงตัวอย่างหน้าจอแสดงผลลัพธ์ที่ได้จากการค้นคืน

จากรูปที่ 3.8 แสดงตัวอย่างหน้าจอแสดงผลลัพธ์ที่ได้จากการค้นคืนสารสนเทศ โดยในส่วนของผลลัพธ์จะแสดงชื่อเอกสารที่ค้นคืนได้ ส่วนของเนื้อหาที่สัมพันธ์กับคำหลักในการค้นหาและที่อยู่ของไฟล์ ในส่วนท้ายของหน้าจะแสดงจำนวนเอกสารที่ค้นคืนได้และเวลาทั้งหมดที่ใช้ในการค้นคืนเอกสาร เมื่อผู้ใช้งานต้องการเปิดเอกสารจากผลลัพธ์ที่แสดง ผู้ใช้งานจะสามารถเลือกคลิกที่ชื่อของเอกสารได้โดยตรง แต่ถ้าผลลัพธ์ที่ได้มาไม่ตรงกับเอกสารที่ผู้ใช้งานต้องการ ผู้ใช้งานสามารถทำการใส่คำหลักในการค้นคืนได้ใหม่ โดยสามารถใส่คำหลักที่ต้องการค้นคืนได้ในช่องใส่คำหลักที่อยู่ด้านบนของหน้าแสดงผลลัพธ์และคลิกที่ปุ่ม Search ระบบจะทำงานเหมือนกับการค้นคืนข้อมูลในรูปที่ 3.7 ทุกประการ อีกส่วนคือคำแนะนำในการค้นคืน ซึ่งระบบจะแนะนำคำที่มีความหมายคล้ายคลึงกับคำหลักที่ผู้ค้นคืนได้ทำการค้นคืนเอกสารแต่ผลลัพธ์ที่ได้อาจไม่ตรงตามที่ผู้ใช้งานต้องการ ผู้ใช้งานสามารถคลิกที่คำที่ระบบแนะนำ ระบบจะทำการค้นคืนผลลัพธ์จากคำแนะนำและแสดงผลลัพธ์ให้ผู้ใช้งานอีกครั้งหนึ่ง ในหน้าแสดงผลลัพธ์นี้ผู้ดูแลระบบยังสามารถที่จะเข้าสู่ระบบได้โดยการคลิกที่ปุ่มเข้าสู่ระบบ โดยระบบจะให้ทำการกรอกชื่อผู้ใช้งานและรหัสผ่านในการเข้าสู่ระบบเพื่อทำการปรับปรุงข้อมูลของระบบ

2. รูปแบบหน้าจอแสดงผลของผู้ดูแลระบบ (Administrator)

ในส่วนของผู้ดูแลระบบ หน้าจอแสดงผลจะมีความแตกต่างจากผู้ดูแลระบบสูงสุดเพียงเล็กน้อย คือระบบจะไม่แสดงเมนู เพิ่มผู้ดูแลระบบและรายชื่อผู้ดูแลระบบ แต่ในส่วนอื่นๆ รูปแบบหน้าจอแสดงผลจะเหมือนกับหน้าจอแสดงผลของผู้ดูแลระบบสูงสุดและมีการทำงานของเมนูต่างๆ เหมือนกันทุกประการ โดยหน้าจอแสดงผลของผู้ดูแลระบบเมื่อเข้าสู่ระบบมาจะมีลักษณะดังแสดงในรูปที่ 3.9



รูปที่ 3.9 แสดงตัวอย่างหน้าจอแสดงผลเมื่อผู้ดูแลระบบเข้าสู่ระบบ

จากรูปที่ 3.9 เป็นหน้าจอแสดงของผู้ดูแลระบบ เมื่อเข้าสู่ระบบแล้วระบบจะแสดงหน้าจอแสดงคำศัพท์ทั้งหมดในระบบ ซึ่งในหน้านี้จะมีการทำงานเหมือนกับหน้าจอแสดงคำศัพท์ทั้งหมดในระบบของผู้ดูแลระบบสูงสุดทุกประการ ในแถบเมนูที่แสดงนั้นสำหรับผู้ดูแลระบบจะไม่มีเมนูเพิ่มผู้ดูแลระบบและรายชื่อผู้ดูแลระบบ นอกนั้นในส่วนอื่นมีการทำงานเหมือนกับหน้าจอของผู้ดูแลระบบสูงสุดทุกประการซึ่งผู้วิจัยจะอธิบายรายละเอียดในหัวข้อที่ 3 ต่อไป

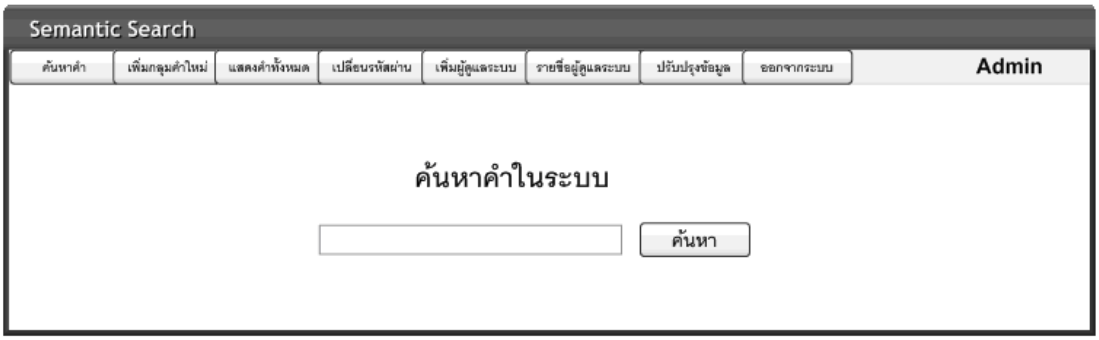
3.รูปแบบหน้าจอแสดงผลของผู้ดูแลระบบสูงสุด (Super Administrator)

ในส่วนของการออกแบบหน้าจอของผู้ดูแลระบบสูงสุดนั้น เนื่องจากการทำงานของระบบมีส่วนที่เกี่ยวข้องกับการค้นคืนสารสนเทศ ดังนั้นผู้ดูแลระบบสูงสุดจะมีหน้าที่ในการแก้ไขปรับปรุงข้อมูลของคำศัพท์ต่างๆ ที่ใช้ในระบบ อีกทั้งยังต้องมีการปรับปรุงดัชนีสำหรับการค้นคืน

สารสนเทศให้มีความทันสมัยอยู่ตลอดเวลารวมถึงการจัดการเพิ่มหรือลดผู้ดูแลระบบอีกด้วย ทำให้หน้าจอของผู้ดูแลระบบสูงสุดถูกออกแบบมาให้มีหน้าจอแสดงผลสำหรับการจัดการต่างๆครบถ้วน โดยผู้ดูแลระบบสูงสุดสามารถเข้าสู่ระบบได้โดยการคลิกปุ่มเข้าสู่ระบบ (Login) จากนั้นกรอกชื่อผู้ใช้งานและรหัสผ่านเมื่อระบบทำการตรวจสอบแล้วจะแสดงหน้าจอแรกดังแสดงในรูปที่ 3.10

รูปที่ 3.10 แสดงตัวอย่างหน้าจอแสดงผลเมื่อผู้ดูแลระบบสูงสุดเข้าสู่ระบบ

ในหน้าแรกหลังจากที่ที่ผู้ดูแลระบบสูงสุดเข้าสู่ระบบมาแล้ว จะแสดงหน้าจอดังแสดงในรูปที่ 3.10 โดยในหน้านี้จะมีแถบการจัดการต่างๆ คือ ค้นหาคำ เพิ่มกลุ่มคำใหม่ แสดงคำทั้งหมด เปลี่ยนรหัสผ่าน เพิ่มผู้ดูแลระบบ ดูรายชื่อผู้ดูแลระบบ ปรับปรุงข้อมูลและออกจากระบบ ซึ่งหน้าแรกที่เข้ามาระบบจะแสดงหน้าคำศัพท์ทั้งหมดในระบบ โดยผู้ดูแลระบบสามารถที่จะทำการปรับปรุงแก้ไขคำศัพท์ได้หรือจะลบคำศัพท์ในระบบทิ้งได้ โดยการคลิกที่ปุ่ม Edit สำหรับแก้ไขคำหรือคลิกที่ปุ่ม Delete สำหรับการลบคำนั้นๆ ออกจากระบบ ในส่วนล่างของหน้าจอนี้จะมีเมนูให้เลือกการแสดงผลจำนวนคำต่อหนึ่งหน้าแสดงผล ในรูปต่อไปผู้วิจัยจะทำการแสดงหน้าจอเมื่อเลือกค้นหาคำในระบบดังแสดงในรูปที่ 3.11



รูปที่ 3.11 แสดงตัวอย่างหน้าค้นหาคำในระบบ

เนื่องจากคำในระบบมีเป็นจำนวนมากและอาจเพิ่มมากขึ้นได้ทุกวัน ทำให้บางครั้งในการเพิ่มคำต้องมีการตรวจสอบว่าในระบบมีคำคำนั้นหรือไม่ หรืออาจจะเป็นการค้นหาคำเพื่อปรับปรุงการเขียนให้ถูกต้องหรือเป็นการค้นหาคำเพื่อลบคำคำนั้นออกจากระบบ โดยผู้ดูแลระบบสูงสุดสามารถค้นหาคำที่ต้องการได้โดยเข้ามาที่หน้าค้นหาคำในระบบนี้ ในหน้าจอนี้ผู้ดูแลระบบสูงสุดสามารถค้นหาคำที่มีอยู่ในระบบได้โดยระบุคำที่ต้องการค้นหาลงในช่องว่าง หลังจากนั้นคลิกที่ปุ่มค้นหา โดยระบบจะทำการค้นหาคำและแสดงผลออกมา โดยมีหน้าจอแสดงผลการค้นหาดังแสดงในรูปที่ 3.12

รูปที่ 3.12 แสดงตัวอย่างหน้าแสดงผลลัพธ์ในการค้นหา

ในส่วนของหน้าจอแสดงผลลัพธ์ในการค้นหา ระบบจะทำการแสดงคำที่มีความสัมพันธ์กับคำที่ค้นหาทั้งหมด ผู้ดูแลระบบสามารถที่จะแก้ไขคำหรือลบคำได้โดยคลิกที่ปุ่ม Edit เพื่อทำการแก้ไขคำ หรือคลิกที่ปุ่ม Delete เพื่อลบคำนั้นๆ อีกส่วนหนึ่งสำหรับหน้าจอแสดงผลนี้ก็คือ การเพิ่มคำที่มีความความหมายคล้ายคลึงกับคำผลลัพธ์ โดยสามารถเพิ่มคำได้ในช่องด้านล่าง ระบบจะทำการตรวจสอบว่ามีคำนี้อยู่ในระบบหรือไม่ ถ้าตรวจสอบไม่พบคำนี้อยู่ในระบบก็จะทำการเพิ่มคำใหม่เข้าสู่ระบบ ในแถบเมนูต่อไปก็คือเมนูเพิ่มกลุ่มคำใหม่ เมื่อเลือกที่เมนูเพิ่มกลุ่มคำใหม่หน้าจอจะแสดงดังรูปที่ 3.13

รูปที่ 3.13 แสดงตัวอย่างหน้าเพิ่มกลุ่มความสัมพันธ์ใหม่

ในส่วนของหน้าเพิ่มกลุ่มความสัมพันธ์ใหม่ หน้านี้จะมีช่องให้เพิ่มกลุ่มความสัมพันธ์ใหม่ได้สูงสุด 5 คำ โดยสามารถเพิ่มได้ตั้งแต่ 1 คำ แต่คำที่เพิ่มในหน้านี้ทั้งหมดต้องเป็นคำที่มีความหมายคล้ายคลึงกันเท่านั้น เช่นคำว่า คอมพิวเตอร์ส่วนบุคคลและพีซี เป็นต้น เมื่อเพิ่มคำเรียบร้อยแล้วผู้ดูแลระบบสามารถคลิกที่ปุ่มบันทึกเพื่อทำการเพิ่มคำได้เลย ถ้าต้องการเปลี่ยนแปลงคำผู้ดูแลระบบสามารถคลิกที่ปุ่มล้างเพื่อล้างคำที่พิมพ์ไปก่อนหน้านี้ได้

ในแถบเมนูต่อไปก็คือเมนูเปลี่ยนรหัสผ่าน เมื่อเลือกที่เมนูเปลี่ยนรหัสผ่าน หน้าจอจะแสดงดังรูปที่ 3.14

รูปที่ 3.14 แสดงตัวอย่างหน้าเปลี่ยนรหัสผ่าน

ในหน้าเปลี่ยนรหัสผ่าน จะเป็นส่วนที่มีไว้เพื่อให้ผู้ดูแลระบบสามารถเปลี่ยนรหัสผ่านในการเข้าสู่ระบบ ในหน้านี้จะมีช่องให้กรอกสามช่อง คือช่องรหัสผ่านเดิม รหัสผ่านใหม่ และยืนยันรหัสผ่านใหม่ เมื่อผู้ดูแลระบบได้ทำการกรอบครบถ้วนทุกช่องแล้วให้คลิกที่ปุ่ม บันทึก ระบบจะทำการเปลี่ยนรหัสผ่านเพื่อเข้าสู่ระบบให้ทันที

ในส่วนถัดไปจะเป็นเมนูเพิ่มผู้ดูแลระบบ เมื่อเลือกที่เมนูเพิ่มผู้ดูแลระบบ ระบบจะแสดงหน้าจอดังรูปที่ 3.15

รูปที่ 3.15 แสดงตัวอย่างหน้าเพิ่มผู้ดูแลระบบ

จากรูปที่ 3.15 แสดงหน้าจอเพิ่มผู้ดูแลระบบ ผู้ที่จะสามารถมองเห็นเมนูนี้และเข้าใช้งานได้ต้องเป็นผู้ดูแลระบบสูงสุดเท่านั้น ในหน้านี้จะมีช่องให้ระบุรหัสผู้ใช้งานที่ต้องการเพิ่ม ช่องระบุรหัสผ่านและช่องยืนยันรหัสผ่าน เมื่อกรอกข้อมูลถูกต้องแล้ว คลิกที่ปุ่มบันทึกหลังจากนั้นระบบจะทำการเพิ่มผู้ดูแลระบบใหม่ทันที ในส่วนถัดไปคือเมนูรายชื่อผู้ดูแลระบบ จะเป็นหน้าจอที่แสดงรายชื่อผู้ดูแลระบบทั้งหมดที่มี ดังแสดงในรูปที่ 3.16

รูปที่ 3.16 แสดงตัวอย่างหน้าแสดงรายชื่อผู้ดูแลระบบ

จากรูปที่ 3.16 คือหน้าจอแสดงรายชื่อผู้ดูแลระบบ ซึ่งระบบจะแสดงชื่อผู้ใช้งานและขั้นของผู้ใช้งานในระบบนี้จะมีอยู่สองขั้นคือ ผู้ดูแลระบบสูงสุด (Super Admin) และ ผู้ดูแลระบบทั่วไป (Admin) อีกส่วนหนึ่งคือช่องปรับปรุงข้อมูล ซึ่งเมื่อเลือกที่ปุ่ม Edit ระบบจะสามารถปรับเปลี่ยนขั้นของผู้ใช้งานได้และอีกปุ่มหนึ่งก็คือปุ่ม Delete ระบบจะทำการลบผู้ดูแลระบบนั้นออกจากระบบ ซึ่งในส่วนนี้จะสามารถเข้าใช้งานได้เฉพาะผู้ดูแลระบบสูงสุดเท่านั้น

ในส่วนถัดไปคือเมนูรายปรับปรุงข้อมูล จะเป็นหน้าจอที่ใช้สำหรับกำหนดไดเรคทอรีของระบบที่จะนำเอกสารในไดเรคทอรีนั้นๆ ไปสร้างดัชนีสำหรับการค้นคืนสารสนเทศ ดังแสดงในรูปที่ 3.17

รูปที่ 3.17 แสดงตัวอย่างหน้าปรับปรุงข้อมูล

จากรูปที่ 3.17 คือหน้าจอปรับปรุงข้อมูล ซึ่งในหน้าจอนี้จะมีช่องให้ผู้ใช้งานได้กำหนดไดเรคทอรีสำหรับสร้างดัชนีสำหรับการค้นคืนสารสนเทศของระบบ โดยระบบจะมีค่าที่เป็นมาตรฐานไว้ให้ ถ้าผู้ดูแลระบบไม่ต้องการเปลี่ยนไดเรคทอรีของเอกสาร ผู้ดูแลระบบสามารถปรับปรุงเอกสารได้โดยเลือกคลิกที่ปุ่ม ปรับปรุงข้อมูลได้ทันที แต่ถ้าผู้ดูแลระบบต้องการปรับเปลี่ยนไดเรคทอรีในการสร้างดัชนีของระบบค้นคืนสารสนเทศ ผู้ดูแลระบบต้องระบุไดเรคทอรีในช่องข้อความให้ถูกต้องหลังจากนั้นให้เลือกที่ปุ่มปรับปรุงข้อมูล ระบบจะทำการปรับปรุงข้อมูลเอกสารของระบบต่อไป

เมนูสุดท้ายคือออกจากระบบ เมื่อผู้ดูแลระบบจัดการปรับปรุงแก้ไขระบบเสร็จสิ้นแล้ว เมื่อเลือกเมนูออกจากระบบ ระบบจะทำการกลับมาที่หน้าแรกของระบบ

3.7 แนวทางและวิธีในการพัฒนาระบบ

ระบบการค้นคืนข้อมูลด้านเทคโนโลยีสารสนเทศภาษาไทย ในขั้นตอนการพัฒนาผู้วิจัยได้จัดเตรียมในส่วนของข้อมูลและซอฟต์แวร์ที่สำคัญ รวมถึงในส่วนของฮาร์ดแวร์ที่ผู้วิจัยได้นำมาใช้ในการพัฒนาระบบค้นคืนข้อมูลเทคโนโลยีสารสนเทศภาษาไทย โดยได้ทำการคัดเลือกมาจากการศึกษาค้นคว้าข้อมูลดังนี้

1. ภาษาหลักในการพัฒนาระบบผู้วิจัยได้เลือกใช้ภาษา Java ในการพัฒนาในส่วนประมวลผลต่าง ๆ และส่วนของการแสดงผลในรูปแบบเว็บแอพพลิเคชันได้ใช้ภาษา JSP (Java Server Page) ซึ่งพัฒนามาจากภาษา Java ทำให้สามารถใช้งานร่วมกันในส่วนแสดงผลและส่วนประมวลผลได้เป็นอย่างดี อีกทั้งภาษา Java มีข้อดีคือสามารถทำงานได้ในทุกระบบปฏิบัติการจึงสามารถลดข้อจำกัดในการใช้งานได้ ในงานวิจัยนี้ผู้วิจัยได้เลือกใช้ Eclipse Java EE IDE Helios เป็นเครื่องมือในการพัฒนาและประมวลผลคำสั่งภาษา Java

2. ใช้ส่วนเสริม Lucene 3.0 ในการพัฒนาส่วนของการพัฒนาระบบค้นคืน เช่นการสร้างดัชนี การจัดการเกี่ยวกับการค้นคืน เป็นต้น Lucene 3.0 ถูกพัฒนาขึ้นมาจากพื้นฐานภาษา Java แต่ในปัจจุบันได้พัฒนาให้สามารถใชงานได้ในหลายภาษา Lucene 3.0 ถูกพัฒนาขึ้นมาเพื่อช่วยในการพัฒนาระบบค้นคืนข้อมูลโดยเฉพาะ โดยมีคลาสที่พร้อมสำหรับการใช้งานในการพัฒนาระบบค้นคืนข้อมูลอย่างครบถ้วน สถาปัตยกรรมเชิงตรรกะของลูซีนคือแนวคิดว่าเอกสารประกอบไปด้วยเขตข้อมูลของข้อความ ซึ่งทำให้ส่วนต่อประสานโปรแกรมประยุกต์ของลูซีนยืดหยุ่นพอที่จะไม่ขึ้นอยู่กับรูปแบบไฟล์ ข้อความจากไฟล์ในรูปแบบ PDF HTML เอกสารไมโครซอฟต์เวิร์ดและรูปแบบอื่นๆ อีกมากมายสามารถนำมาสร้างดัชนีได้ตราบเท่าที่สามารถสกัดข้อความจากเอกสารได้ อีกทั้งยังมีคลาสที่ใช้สำหรับจัดการในส่วนของการแสดงผลลัพธ์อีกด้วย โดยส่วนเสริม Lucene 3.0 นี้เป็นซอฟต์แวร์ฟรี สามารถนำมาใช้งานได้โดยไม่มีปัญหาด้านลิขสิทธิ์

3. เลือกใช้งาน Thai Wordnet ซึ่งเป็นออนโทโลจีตัวหนึ่ง ซึ่งรวบรวมกลุ่มความสัมพันธ์ของคำภาษาไทยในรูปแบบเชิงความหายเอาไว้จำนวนเกือบหนึ่งแสนคำ ซึ่งสามารถดาวน์โหลดเพื่อนำมาใช้งานได้โดยไม่เสียค่าใช้จ่าย โดยในการนำมาใช้งานผู้วิจัยได้ทำการนำมาพัฒนาเพื่อให้สามารถใช้งานร่วมกับภาษาจาวาและ Lucene 3.0 โดยแปลงข้อมูลต้นฉบับเพื่อให้เหมาะกับการนำมาใช้งานในรูปแบบเว็บ ซึ่งผู้วิจัยจะไม่มีการนำระบบฐานข้อมูลเข้ามาใช้จำต้องแปลงให้อยู่ในรูปแบบของ XML file อีกทั้งจัดการความสัมพันธ์ของคำให้เหมาะสมกับการใช้งานด้านเทคโนโลยีสารสนเทศภาษาไทยมากขึ้น

4. ใช้ส่วนเสริม ThaiAnalyzer 3.0 ซึ่งพัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) เป็นพื้นฐานในการพัฒนาการตัดคำภาษาไทย โดยส่วนเสริมนี้มีพื้นฐานในการตัดคำภาษาไทยโดยหลักการตัดคำยาวที่สุดซึ่งมีความถูกต้องสูง ในการนำมาใช้งานผู้วิจัยจะใช้ ThaiAnalyzer 3.0 ในส่วนของการตัดคำหลักในการค้นคืน เพื่อคัดเลือกคำสำคัญที่จะในไปใช้งาน ในอีกส่วนหนึ่งคือการตัดคำเพื่อใช้ในการสร้างดัชนี เพื่อคัดคำที่สำคัญในการจัดทำดัชนี

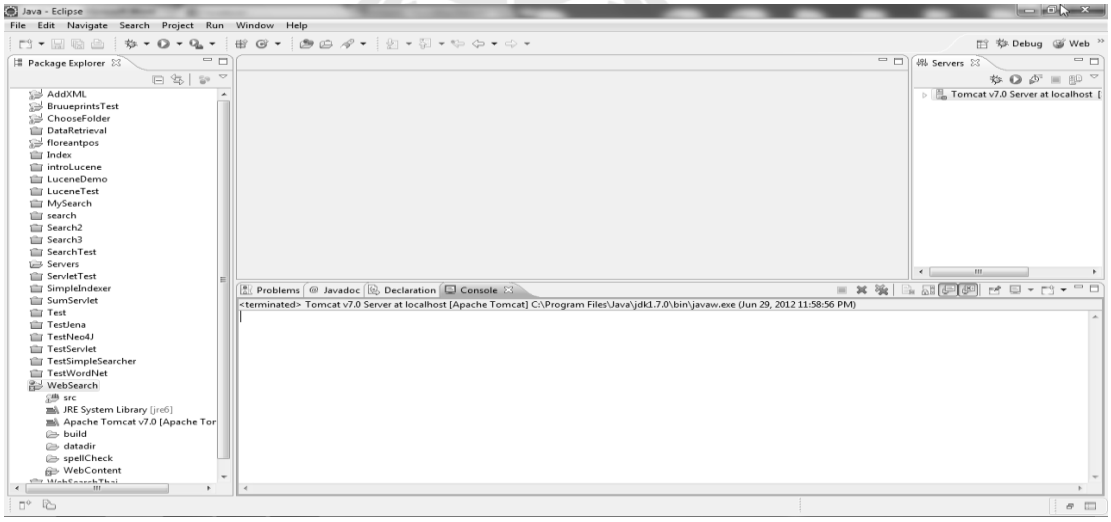
อีกทั้งผู้วิจัยยังได้ทำการพัฒนาต่อยอดในการตัดคำภาษาไทย เพื่อใช้ในการค้นคืนด้านเทคโนโลยีสารสนเทศภาษาไทย โดยพัฒนาเพิ่มคลังข้อความในส่วนของคำที่เขียนตามเสียงภาษาอังกฤษที่เกี่ยวข้องด้านเทคโนโลยีสารสนเทศ และเพิ่มคำสั่งในการตัดคำโดยเลือกคำที่สั้นที่สุดอีกด้วย

5. ใช้ Apache Tomcat 7.0 เป็น Web Server ในการประมวลผลของฝั่งเซิร์ฟเวอร์ เนื่องจากสามารถทำงานร่วมเว็บแอพพลิเคชันที่พัฒนาด้วยภาษา Java ได้เป็นอย่างดี มีความยืดหยุ่นในการใช้งานและยังมีส่วนเสริมต่างๆ ที่ช่วยเพิ่มความสะดวกในการใช้งาน

เมื่อได้จัดเตรียมทรัพยากรที่จำเป็นเรียบร้อยแล้ว ในขั้นตอนต่อไปผู้วิจัยจะได้อธิบายถึงขั้นตอนในการพัฒนาระบบค้นคืนข้อมูลเทคโนโลยีสารสนเทศภาษาไทยต่อไป

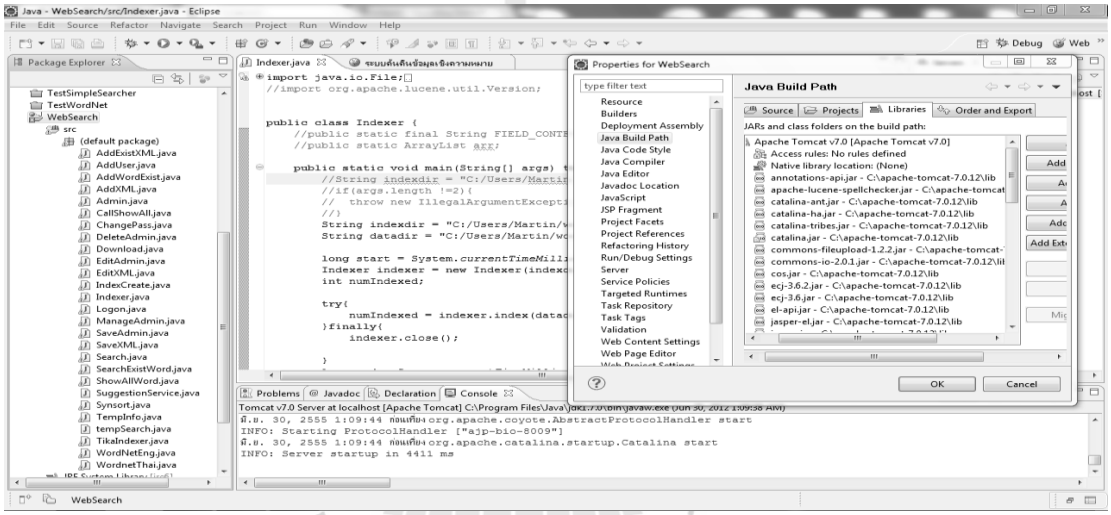
3.7.1 ขั้นตอนการติดตั้งเครื่องมือในการพัฒนาระบบ

ในส่วนของการติดตั้งเครื่องมือในการพัฒนาระบบผู้วิจัยได้ทำการดาวน์โหลด Eclipse Java EE IDE ซึ่งสามารถดาว์นโหลดได้จากเว็บไซต์ heep://www.eclipse.org ซึ่งสามารถนำมาใช้งานได้โดยไม่เสียค่าใช้จ่าย จากนั้นทำการติดตั้งในเครื่องคอมพิวเตอร์ซึ่งมีระบบปฏิบัติการ Windows 7 Ultimate 32-bits ซึ่งเมื่อทำการติดตั้งสำเร็จแล้วเมื่อเข้าสู่หน้าของโปรแกรมจะมีลักษณะดังแสดงในรูปที่ 3.18



รูปที่ 3.18 แสดงหน้าจอของ Java Eclipse EE IDE Helios

เมื่อติดตั้ง Java Eclipse EE IDE เรียบร้อยแล้วจากนั้นผู้วิจัยจะทำการสร้าง Project ขึ้นมาและทำการติดตั้งส่วนเสริมต่างๆ ที่จำเป็นต่อการพัฒนาระบบ โดยส่วนเสริมที่สำคัญได้แก่ Lucene 3.0 Apache Tomcat 7.0 และ Java Servlet โดยเมื่อทำการติดตั้งเรียบร้อยแล้วระบบก็สามารถที่จะใช้งานในการพัฒนาเว็บแอพพลิเคชันได้ โดยเมื่อทำการพัฒนาระบบอาจมีการติดตั้งส่วนเสริมเพิ่มเติมเพื่อความสะดวกในการพัฒนา โดยส่วนเสริมหลักๆ และหน้าจอรูปแบบการพัฒนาระบบที่ผู้วิจัยได้ทำการพัฒนามีรูปแบบดังแสดงในรูปที่ 3.19



รูปที่ 3.19 แสดงหน้าจอของการพัฒนาระบบ

3.7.2 ขั้นตอนการพัฒนาระบบ

ในขั้นตอนการพัฒนาระบบ เมื่อได้ทำการเตรียมเครื่องมือต่างๆ เรียบร้อยแล้ว ในส่วนของขั้นตอนแรกจะทำการพัฒนาโปรแกรมในส่วนของการสร้างดัชนีของข้อมูล โดยการพัฒนาจะทำโดยใช้ภาษา Java ร่วมกับส่วนเสริม Lucene 3.0 โดยการใช้งานในส่วนนี้จะเป็นรูปแบบการเขียนเช่นเดียวกับภาษา Java เพียงแต่การใช้งานส่วนใหญ่จะอ้างอิงคุณสมบัติมาจาก Lucene 3.0 ดังมีตัวอย่างคำสั่งที่ใช้ในการสร้างดัชนีดังแสดงในรูปที่ 3.20

```
try {
            parser.parse(is, handler, metadata,new ParseContext());
            } finally {
                is.close();
            }

            Document doc = new Document();
            doc.add(new Field("contents", handler.toString(),Field.Store.YES, Field.Index.ANALYZED));
            if (DEBUG) {
                        System.out.println(" all text: " + handler.toString());
                        }
            for(String name : metadata.names()) {
                String value = metadata.get(name);
                if (textualMetadataFields.contains(name)) {
            doc.add(new Field("contents", value,Field.Store.YES, Field.Index.NOT_ANALYZED));
                    }
            doc.add(new Field(name, value,Field.Store.YES, Field.Index.NO));
            if (DEBUG) {
                System.out.println(" " + name + ": " + value);
                    }
                }
            if (DEBUG) {
                System.out.println();
                }
            doc.add(new Field("filename", f.getCanonicalPath(),Field.Store.YES, Field.Index.NOT_ANALYZED));
            return doc;
            }
        }
```

รูปที่ 3.20 แสดงตัวอย่างคำสั่งภาษา Java ในการสร้างดัชนีจากเอกสารที่ต้องการ

เมื่อได้ทำการสร้างในส่วนของการสร้างดัชนีแล้วขั้นตอนต่อมาจะเป็นการพัฒนาในส่วนของการค้นคืนข้อมูล โดยการพัฒนาจะใช้เครื่องมือในการพัฒนาเช่นเดียวกับการพัฒนาการสร้างดัชนี ดังมีตัวอย่างที่ใช้ในการพัฒนาส่วนค้นคืนข้อมูลดังแสดงในรูปที่ 3.21

```
Document doc = is.doc(scoreDoc.doc);
Formatter formatter = new SimpleHTMLFormatter("<font color=red><b>","</b></font>");
QueryScorer scorer = new QueryScorer(query,"contents");
Encoder encoder = new SimpleHTMLEncoder();
Highlighter highlighter = new Highlighter(formatter,encoder,scorer);
highlighter.setTextFragmenter(new SimpleFragmenter(200));

    result = highlighter.getBestFragment(
            new ThaiAnalyzer(), "contents", doc.get("contents"));
```

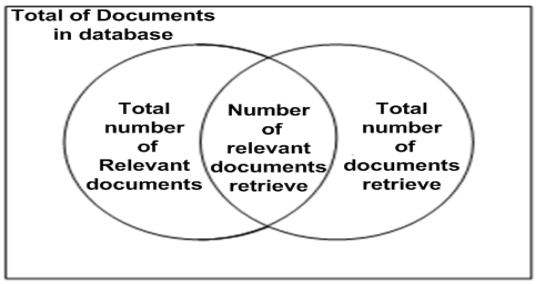
รูปที่ 3.21 แสดงตัวอย่างคำสั่งภาษา Java ในการค้นคืนข้อมูลจากดัชนีที่สร้างขึ้น

จากรูปที่ 3.21 จะเห็นว่ามีการค้นคืนข้อมูลโดยค้นคืนข้อมูลจาก Contents ของเอกสาร และในส่วนของการแสดงผลผู้วิจัยได้ใช้คลาส Formatter เพื่อแสดงผลในหน้าแสดงผลให้คำหลักที่ใช้ค้นคืนมีสีแดง ซึ่งจะทำให้ผู้ใช้งานสามารถสังเกตเห็นได้โดยง่าย

3.8 แนวทางและวิธีการทดสอบระบบ

3.8.1 แนวทางการทดสอบโดยการประเมินผลด้านความแม่นยำและความถูกต้อง

จากแนวคิดในการค้นคืนข้อมูลเชิงความหมายด้านเทคโนโลยีสารสนเทศภาษาไทยที่ผู้วิจัยได้นำเสนอในตอนต้นนั้น สามารถทำการทดสอบและประเมินผลได้โดยใช้การวัดค่าความแม่นยำ (Precision) ค่าความถูกต้อง (Recall) โดยจะแสดงดังรูปที่ 3.22



รูปที่ 3.22 แสดงเซตของเอกสารทั้งหมดที่ใช้ในการทดสอบ

สมการที่ใช้หาค่าประสิทธิภาพ Precision และ Recall แสดงได้ดังนี้คือ

Precision = Number of relevant documents retrieve / Total number of documents retrieve (3-1)

จากสมการที่ 3-1 คำนวณค่าความแม่นยำได้จากจำนวนผลลัพธ์ที่เกี่ยวข้องและถูกค้นคืนทั้งหมดหารด้วยจำนวนเอกสารทั้งหมดเป็นผลลัพธ์ของการค้นคืน

Recall = Number of relevant documents retrieve / Total number of relevant documents

(3-2)

จากสมการที่ 3-2 คำนวณค่าความแม่นยำได้จากจำนวนผลลัพธ์ที่เกี่ยวข้องและถูกค้นคืนทั้งหมดหารด้วยจำนวนเอกสารทั้งหมดที่เกี่ยวข้อง

ในการทดสอบและประเมินผลความแม่นยำผู้วิจัยจะใช้เอกสารที่เกี่ยวข้องจำนวน 50 ชุด ซึ่งเอกสารทั้งหมดเป็นเอกสารภาษาไทยที่อยู่ในรูปแบบข้อมูล PDF File Format และผู้วิจัยจะทำการสุ่มทดลองค้นคืนข้อมูลโดยใช้คำหลักต่างกัน 20 คำ ซึ่งเป็นคำภาษาไทยที่เกี่ยวข้องกับด้านเทคโนโลยีสารสนเทศ ผลลัทธ์ที่ได้จะถูกคำนวณค่าความถูกต้องตามสมการที่ 3-1 และ 3-2 ในส่วนของเกณฑ์ในการประเมินประสิทธิภาพของระบบคือค่าความแม่นยำ (Precision) และค่าความถูกต้อง (Recall) ต้องมีค่ามากกว่า 60 % จึงจะถือว่าระบบที่พัฒนามีประสิทธิภาพดี

3.8.2 แนวทางในการประเมินผลโดยผู้ใช้งาน

ในส่วนของการประเมินโดยผู้ใช้งาน ผู้วิจัยจะได้จัดทำแบบสำรวจความคิดเห็นและการให้คะแนนในการทดสอบระบบ โดยผู้ที่จะเข้าร่วมการทดสอบจะแบ่งเป็นสองกลุ่มคือ กลุ่มนักวิจัยหรือนักพัฒนาที่เกี่ยวข้องกับสายงานด้านเทคโนโลยีสารสนเทศและกลุ่มผู้ใช้งานทั่วไป

ในบทต่อไป ผู้วิจัยจะแสดงถึงข้อสรุปของผลการทดลองและการอภิปรายผลการทดลอง โดยจะมีรายละเอียดที่สำคัญคือ สรุปผลการทดลองในส่วนของประสิทธิภาพความรวดเร็วในการประมวลผลของระบบ สรุปผลการทดลองในส่วนของความถูกต้องและแม่นยำและสรุปและอภิปรายผลการทดลองโดยรวม

บทที่ 4
การพัฒนาและการทดสอบระบบ

การจัดทำระบบค้นคืนสารสนเทศด้านเทคโนโลยีสารสนเทศภาษาไทยโดยใช้หลักการเชิงความหมาย ได้นำทฤษฎีและหลักการของเว็บแอพพลิเคชันซึ่งพัฒนาโดยมีพื้นฐานเป็นภาษาจาวา ซึ่งเป็นภาษาที่สามารถทำงานร่วมกับหลักการเชิงความหมายได้เป็นอย่างดี ในการทดสอบและอภิปรายผลผู้วิจัยจะได้จำแนกรายละเอียดเป็นหัวข้อดังนี้คือ สภาพแวดล้อมที่ใช้ในการพัฒนาและทดสอบระบบ โครงสร้างของระบบ การทดสอบระบบและการอภิปรายผล

4.1 สภาพแวดล้อมที่ใช้ในการพัฒนาและทดสอบระบบ

เครื่องคอมพิวเตอร์ที่ใช้ในการพัฒนาระบบมีรายละเอียดดังนี้

ฮาร์ดแวร์

- คอมพิวเตอร์โน๊ตบุค Core 2 Duo 2.2 กิกะเฮิร์ส
- หน่วยความจำภายใน 2.7 กิกะไบต์
- ฮาร์ดดิสก์ 160 กิกะไบต์
- หน้าจอแสดงผล 13 นิ้ว

ซอฟต์แวร์

- ระบบปฏิบัติการ Windows7 Ultimate 32bits
- ระบบปฏิบัติการ Mac OSX 10.6
- Eclipse Java EE IDE Helios
- Apache Tomcat 7.0
- Java Servlet API
- Adobe Photoshop CS3
- Google Chrome Web Browser Version 20.0.1132.57
- JRE 1.7 (Java Runtime Environment 1.7)

คำศัพท์ที่มีในระบบ

- คำศัพท์ภาษาอังกฤษจากฐานข้อมูลเวิร์ดเน็ต จำนวน 207788 คำ
- คำศัพท์ภาษาไทยจากฐานข้อมูลคำไทยจากศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) จำนวน 4000 คำ
- คำศัพท์ด้านเทคโนโลยีสารสนเทศซึ่งผู้วิจัยได้คัดกรองมา จำนวน 300 คำ

4.2 โครงสร้างของระบบ

โครงสร้างของระบบที่ผู้วิจัยได้ออกแบบและพัฒนาขึ้นประกอบด้วย 3 ส่วน คือ

1. ส่วนของผู้ใช้งาน ผู้ใช้งานจะสามารถทำได้เฉพาะการค้นคืนสารสนเทศเท่านั้น ไม่สามารถเข้าสู่ระบบเพื่อปรับปรุงแก้ไขระบบได้ โดยหน้าจอของผู้ใช้งานแสดงดังรูปที่ 4.1 ซึ่งหน้าจอของผู้ใช้งานจะมีเพียงช่องเพื่อใส่คำหลักในการค้นคืนสารสนเทศและปุ่มค้นหาเท่านั้น

รูปที่ 4.1 แสดงหน้าจอสำหรับผู้ใช้งาน

2. ส่วนของผู้ดูแลระบบ ผู้ดูแลระบบไปจะถูกแต่งตั้งโดยผู้ดูแลระบบสูงสุดเท่านั้น โดยผู้ดูแลระบบจะมีหน้าที่ในการ แก้ไขปรับปรุงคำศัพท์ในระบบและสร้างดัชนีเอกสารสำหรับการค้นคืนสารสนเทศเท่านั้น ผู้ดูแลระบบไม่สามารถเพิ่มหรือลดผู้ดูแลระบบได้ โดยหน้าจอของผู้ดูแลระบบจะแสดงดังรูปที่ 4.2 ซึ่งเมื่อผู้ดูแลระบบเข้าสู่ระบบสำเร็จแล้วหน้าจอแรกของผู้ดูแลระบบคือหน้าจอแสดงคำศัพท์ทั้งหมดเช่นเดียวกับหน้าจอผู้ดูแลระบบสูงสุด แต่ส่วนที่ต่างกันคือเมนูด้านบน

ของหน้าจอ ซึ่งผู้ดูแลระบบจะไม่มีเมนูที่เกี่ยวข้องกับการเพิ่มลดหรือปรับปรุงผู้ดูแลระบบซึ่งเป็นสิทธิ์เฉพาะผู้ดูแลระบบสูงสุดเท่านั้น

รูปที่ 4.2 แสดงหน้าจอเมื่อผู้ดูแลระบบเข้าสู่ระบบ

3. ส่วนของผู้ดูแลระบบสูงสุด ผู้ดูแลระบบสูงสุดจะทำหน้าที่ที่เกี่ยวข้องกับการปรับปรุงแก้ไขข้อมูลทุกอย่างที่เกี่ยวข้องกับระบบคือ ปรับปรุงหรือแก้ไขข้อมูลคำศัพท์ในระบบ ปรับปรุงแก้ไขข้อมูลผู้ใช้งานและการสร้างดัชนีสำหรับการค้นคืนสารสนเทศ โดยหน้าจอผู้ดูแลระบบสูงสุดเมื่อทำการเข้าสู่ระบบแล้วจะแสดงดังรูปที่ 4.3 ซึ่งจากรูปจะเห็นว่าเมื่อผู้ดูแลระบบสูงสุดเข้าสู่ระบบแล้ว หน้าจอของระบบจะแสดงหน้าจอแสดงคำศัพท์ทั้งหมดเป็นหน้าจอแรก โดยผู้แลระบบสูงสุดสามารถจัดการระบบในส่วนต่างๆ ได้โดยการเลือกที่เมนูต่างๆ ที่อยู่ด้านบนของหน้าจอ

รูปที่ 4.3 แสดงหน้าจอเมื่อผู้ดูแลระบบสูงสุดเข้าสู่ระบบ

4.3 การทดสอบระบบ

4.3.1 ข้อมูลที่ใช้ในการทดสอบระบบ

ในการทดสอบระบบ ผู้วิจัยได้จัดเตรียมเอกสารเชิงวิชาการที่เกี่ยวข้องกับด้านเทคโนโลยีสารสนเทศ โดยเอกสารทั้งหมดเป็นภาษาไทยและจะมีรูปแบบของไฟล์ที่นำมาทดสอบคือ PDF Format เอกสารทั้งหมดที่นำมาทดสอบจะมี 50 ชุด ขนาดไฟล์ของแต่ละเอกสารจะมีขนาดตั้งแต่ 1 เมกกะไบต์ ไปจนถึง 5 เมกกะไบต์

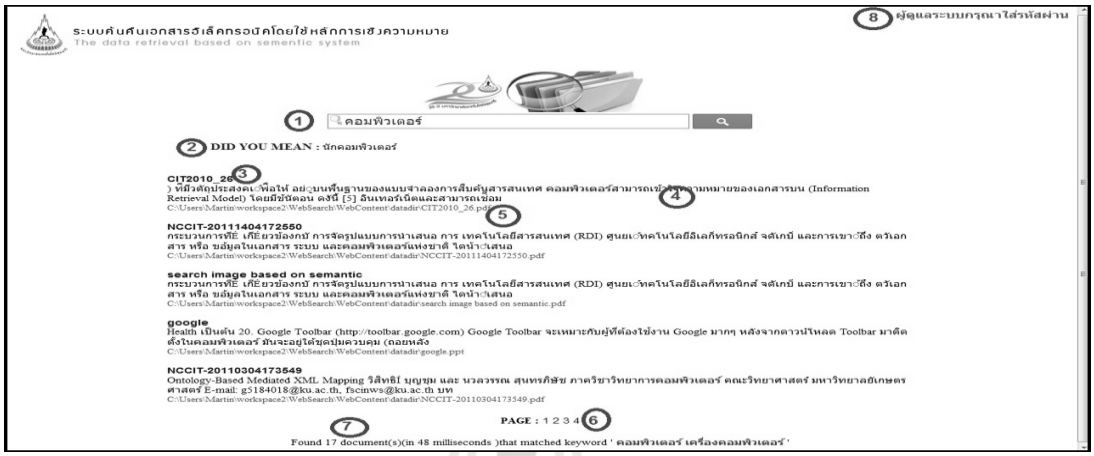
4.3.2 ขั้นตอนการทดสอบระบบ

ในการทดสอบระบบผู้วิจัยจะแบ่งการทดสอบออกเป็น 3 ส่วนด้วยกัน ส่วนแรกคือส่วนของการทดสอบใช้งานของผู้ใช้งาน ส่วนที่สองเป็นส่วนการใช้งานของผู้ดูแลระบบและส่วนที่สามคือส่วนของการทดสอบประสิทธิภาพในการค้นคืนสารสนเทศของระบบ โดยการทดสอบในส่วนของประสิทธิภาพในการค้นคืนสารสนเทศ ผู้วิจัยจะทำการเปรียบเทียบค่าความแม่นยำ (Precision) และค่าการเรียกคืน (Recall) โดยทำการทดสอบเปรียบเทียบกับการค้นคืนสารสนเทศในระบบปฏิบัติการ Windows7 Ultimate 32 bits

1. การทดสอบการใช้งานของผู้ใช้งาน

ผู้ใช้งานจะสามารถใช้ระบบได้เพียงส่วนของการค้นคืนสารสนเทศเท่านั้น โดยเมื่อผู้ใช้งานเข้าสู่ระบบมาจะแสดงหน้าจอดังรูปที่ 4.1 เมื่อผู้ใช้งานต้องการค้นหาเอกสารหรือสารสนเทศใดๆ ให้ผู้ใช้งานระบุคำค้นหาในช่องพิมพ์คำค้นหา โดยระบบจะมีระบบช่วยสะกดคำ

ด้วย เมื่อระบุคำค้นหาเสร็จแล้วให้ผู้ใช้งานคลิกที่ปุ่ม  ระบบจะทำการค้นหาและจะแสดงผลลัพธ์ดังแสดงในรูปที่ 4.4



รูปที่ 4.4 แสดงหน้าจอผลลัพธ์การค้นคืนสารสนเทศ

ในหน้าจอแสดงผลลัพธ์ของการค้นคืนสารสนเทศ จะมีส่วนที่แสดงข้อมูลของแต่ละส่วน ซึ่งผู้วิจัยจะอธิบายตามลำดับของตัวเลขที่แสดงในรูปที่ 4.4

1. ช่องระบุคำค้นหาโดยหน้าแสดงผลจะยังคงค้างคำค้นหาที่ผู้ใช้งานระบุมาจากหน้าแรก
2. ส่วนเสนอแนะ ในส่วนนี้เป็นส่วนที่ระบบจะแนะนำคำที่เขียนหรือมีความหมายคล้ายคลึงกับคำหลักที่ผู้ใช้งานระบุ โดยในกรณีที่ผู้ใช้งานค้นคืนสารสนเทศแล้วผลลัพธ์ที่ได้ไม่ตรงตามที่ต้องการ ผู้ใช้งานอาจเลือกใช้คำค้นหาจากคำที่ระบบแนะนำ โดยคลิกที่คำที่ระบบแนะนำ ระบบจะทำการค้นหาและแสดงผลของคำค้นหาที่แนะนำ
3. ชื่อเอกสาร จากจำนวนเอกสารทั้งหมดที่ระบบค้นคืนได้ แต่ละเอกสารจะมีรายละเอียดแสดงเพื่อให้ผู้ใช้งานตรวจสอบความถูกต้อง โดยเมื่อคลิกที่ชื่อเอกสารระบบจะทำการเปิดเอกสารฉบับเต็มให้ผู้ใช้งานได้ตรวจสอบอีกครั้งหนึ่ง
4. เนื้อหาของเอกสารที่ตรงกับคำหลักในการค้นหา ระบบจะแสดงเนื้อหาในส่วนที่ระบบค้นพบว่ามีความหมายคล้ายคลึงกับคำหลักที่ผู้ใช้งานระบุเพื่อทำการค้นหา
5. ที่อยู่ของเอกสาร ส่วนนี้จะแสดงที่อยู่ของเอกสารว่าอยู่ในไดเรคทอรีใด
6. แสดงจำนวนหน้าทั้งหมดที่ระบบแสดงผลลัพธ์ซึ่งระบบทำการค้นคืนสำเร็จ

7. แสดงจำนวนเอกสารที่ระบบทำการค้นคืนสำเร็จ และเวลาที่ระบบใช้ในการค้นคืนทั้งหมด

8. หากผู้ดูแลระบบต้องการเข้าสู่ระบบ สามารถเข้าสู่ระบบจากหน้าแสดงผลลัพธ์ได้โดยการคลิกที่ข้อความจะมีช่องให้กรอกชื่อผู้ใช้งานและรหัสผ่านเพื่อทำการเข้าสู่ระบบ

2. การทดสอบการใช้งานในส่วนของผู้ดูแลระบบ

ในงานวิจัยนี้ผู้วิจัยได้กำหนดระดับชั้นของผู้ดูแลระบบไว้ 2 ขั้นคือ ผู้ดูแลระบบสูงสุดและผู้ดูแลระบบ ในส่วนนี้ผู้วิจัยจะแสดงการทดสอบระบบในส่วนของผู้ดูแลระบบสูงสุดเนื่องจากเมนูการทำงานของผู้ดูแลระบบทั่วไปนั้นเหมือนเมนูของผู้ดูแลระบบสูงสุดทุกเมนู อีกทั้งเมนูเหล่านั้นมีการทำงานเหมือนกัน สิ่งที่แตกต่างคือเมนูการจัดการสิทธิ์ของผู้ดูแลระบบนั้นผู้ดูแลระบบทั่วไปจะไม่สามารถเข้าใช้งานส่วนนี้ได้ ดังที่ผู้วิจัยได้อธิบายไว้ในบทที่ 3 เมื่อผู้ดูแลระบบเข้าสู่ระบบมาจะพบหน้าแรกดังรูปที่ 4.2 โดยจะเป็นหน้าแสดงคำศัพท์ทั้งหมดในระบบ เมื่อผู้ดูแลระบบต้องการแก้ไขปรับปรุงส่วนหนึ่งส่วนใดของระบบ จะสามารถทำได้โดยกดเลือกที่เมนูในส่วนบนของหน้าจอ ซึ่งจะมีเมนูต่างๆ สำหรับจัดการระบบโดยมีการทำงานแตกต่างกันดังนี้

เมนูค้นหาคำ เมนูนี้เป็นเมนูสำหรับจัดการคำโดยผู้ดูแลระบบสามารถค้นหาคำศัพท์ในระบบ เพื่อทำการปรับปรุงแก้ไขคำศัพท์ได้ โดยสามารถระบุคำที่ต้องการค้นหาในช่องค้นหาคำ ระบบจะทำการค้นหาคำและแสดงผลลัพธ์ในการค้นหาให้ผู้ดูแลระบบตรวจสอบ อีกทั้งยังสามารถปรับปรุงแก้ไขคำที่เป็นผลลัพธ์ได้ในหน้าแสดงผลนี้ด้วยเช่นกัน โดยจะมีหน้าจอแสดงผลดังรูปที่ 4.5

รูปที่ 4.5 แสดงหน้าจอค้นหาคำศัพท์ที่มีในระบบ

เมนูเพิ่มกลุ่มความสัมพันธ์ใหม่ เมนูนี้เป็นเมนูสำหรับเพิ่มกลุ่มความสัมพันธ์ใหม่ โดยสามารถเพิ่มได้ตั้งแต่ 1 ถึง 5 คำต่อการเพิ่มความสัมพันธ์ในแต่ละครั้ง การเพิ่มกลุ่มความสัมพันธ์ใหม่นี้คำที่เพิ่มเข้ามาต้องมีความหมายคล้ายคลึงกัน เช่น คอมพิวเตอร์ พีซี คอมพิวเตอร์ส่วนบุคคล เป็นต้น โดยจะมีหน้าจอแสดงดังรูปที่ 4.6

รูปที่ 4.6 แสดงหน้าจอเพิ่มกลุ่มความสัมพันธ์ใหม่

เมนูแสดงคำทั้งหมด เมนูนี้เป็นเมนูที่ใช้สำหรับดูคำศัพท์ทั้งหมดในระบบ โดยผู้ดูแลระบบสามารถปรับปรุงแก้ไขคำได้ผ่านหน้าเมนูนี้ หน้าจอเมนูแสดงคำทั้งหมดนี้สามารถเลือกจำนวนคำศัพท์ต่อการแสดงผลในหนึ่งหน้าได้ อีกทั้งยังสามารถทำการเรียงคำตามตัวอักษรได้เพื่อให้ง่ายต่อการจัดการคำศัพท์ โดยหน้าจอแสดงคำทั้งหมดจะมีลักษณะดังรูปที่ 4.1

เมนูเปลี่ยนรหัสผ่าน เมนูนี้มีไว้สำหรับเปลี่ยนรหัสผ่านเพื่อเข้าสู่ระบบของผู้ดูแลระบบ โดยจะมีช่องให้กรอกรหัสผ่านเดิม รหัสผ่านใหม่และยืนยันรหัสผ่านใหม่ เมื่อทำการเลือกปุ่มบันทึก ระบบจะทำการเปลี่ยนรหัสผ่านให้ผู้ดูแลระบบทันที โดยหน้าจอเมนูเปลี่ยนรหัสผ่านจะแสดงดังรูปที่ 4.7

รูปที่ 4.7 แสดงหน้าจอเปลี่ยนรหัสผ่าน

เมนูเพิ่มผู้ดูแลระบบ เมนูนี้จะเข้าถึงได้เฉพาะผู้ดูแลระบบสูงสุดเท่านั้น ซึ่งเมนูนี้จะมีไว้เพื่อเพิ่มผู้ดูแลระบบทั่วไป เพื่อให้สามารถเข้ามาแก้ไขปรับปรุงระบบได้ โดยการเพิ่มผู้ดูแลระบบนั้นเมื่อทำการเพิ่มเสร็จแล้วจะยังได้สิทธิ์เพียงแค่ผู้ดูแลระบบทั่วไปเท่านั้น ถ้าต้องการเปลี่ยนสิทธิ์เป็นผู้ดูผลระบบสูงสุดต้องได้รับการเปลี่ยนโดยผู้ดูแลระบบสูงสุดเท่านั้น ในหน้าเมนูนี้จะมีช่องให้กรอกชื่อผู้ใช้งานที่ต้องการเพิ่ม รหัสผ่านและยืนยันรหัสผ่าน เมื่อทำการเลือกที่ปุ่มบันทึกระบบจะทำการเพิ่มผู้ดูแลระบบเข้าสู่ระบบทันที โดยมีหน้าจอแสดงผลดังรูปที่ 4.8

รูปที่ 4.8 แสดงหน้าจอเมนูเพิ่มผู้ดูแลระบบ

เมนูรายชื่อผู้ดูแลระบบ เมนูนี้จะมีไว้เพื่อให้ผู้ดูแลระบบสูงสุดตรวจสอบปรับปรุงแก้ไขผู้ดูแลระบบที่มีอยู่ โดยสามารถที่จะปรับเปลี่ยนสิทธิ์ของผู้ดูแลระบบได้รวมทั้งยังสามารถลบรายชื่อผู้ดูแลระบบออกได้อีกด้วย โดยเมนูนี้มีหน้าจอแสดงผลดังแสดงในรูปที่ 4.9

รูปที่ 4.9 แสดงหน้าจอเมนูรายชื่อผู้ดูแลระบบ

เมนูปรับปรุงข้อมูล เมนูนี้มีไว้เพื่อทำการสร้างดัชนีสำหรับการค้นคืนสารสนเทศ โดยการสร้างดัชนีสำหรับการค้นคืนสารสนเทศนั้น ผู้ดูแลระบบต้องนำเอกสารต้นฉบับไปไว้ในไดเรคทอรีที่กำหนดไว้หลังจากนั้น ทำการกดปุ่มปรับปรุงข้อมูล ระบบจะทำการสร้างดัชนีขึ้นมาใหม่ แต่หากผู้ดูแลระบบต้องการเปลี่ยนไดเรคทอรีที่เก็บไฟล์ต้นฉบับ ผู้ดูแลระบบสามารถทำได้โดยการระบุไดเรคทอรีที่ต้องการให้ระบบสร้างดัชนีจากไดเรคทอรีนั้นในช่องข้อความ หลังจากนั้นกดปุ่มปรับปรุงข้อมูล ระบบจะทำการสร้างดัชนีจากไดเรคทอรีที่ผู้ดูผลระบบระบุ โดยหน้าจอปรับปรุงข้อมูลจะแสดงดังรูปที่ 4.10

รูปที่ 4.10 แสดงหน้าจอเมนูปรับปรุงข้อมูล

เมนูออกจากระบบ เมื่อผู้ดูแลระบบทำการปรับปรุงแก้ไขเสร็จสิ้นแล้ว ผู้ดูแลจะต้องออกจากระบบเพื่อความปลอดภัยของระบบ เนื่องจากถ้าผู้ดูแลระบบไม่ทำการออกจากระบบแล้วมีผู้ไม่หวังดีเข้ามาใช้ต่ออาจทำให้คำศัพท์ในระบบหายหรือผิดพลาดได้ เมื่อผู้ดูแลระบบเลือกที่เมนูออกจากระบบ ระบบจะทำการกลับสู่หน้าค้นหาคำดังรูปที่ 4.3

3. การทดสอบประสิทธิภาพในการค้นคืนสารสนเทศของระบบ

ในงานวิจัยนี้ ผู้วิจัยได้เตรียมการทดสอบประสิทธิภาพในการค้นคืนสารสนเทศของระบบไว้ 3 ส่วนคือการทดสอบโดยการเปรียบเทียบกับการค้นคืนสารสนเทศพื้นฐานของระบบปฏิบัติการ Windows 7 Ultimate 32 bits ในการทดสอบผู้วิจัยจะใช้ข้อมูลชุดเดียวกับในการทดสอบโดยจะทำการทดสอบทั้งสิ้น 10 ชุดการทดสอบ ผู้วิจัยจะทำการเพิ่มชุดข้อมูลในการทดสอบในการทดสอบแต่ละครั้ง โดยครั้งแรกผู้วิจัยจะเริ่มทำการทดสอบที่ข้อมูลทดสอบจำนวน 5 ชุด หลังจากนั้นจะเพิ่มข้อมูลทดสอบครั้งละ 10 ชุด ไปจนถึงการทดสอบครั้งสุดท้ายจะมีข้อมูลทดสอบทั้งสิ้น 50 ข้อมูล ในการทดสอบส่วนนี้ผู้วิจัยจะทำการเปรียบเทียบผลลัพธ์ที่ได้จากการค้นคืนสารสนเทศว่ามีการแสดงผลลัพธ์ในการค้นคืนสารสนเทศทั้งหมดในการค้นแต่ละครั้งจำนวนเท่าไร และเนื้อหาของผลลัพธ์ที่ได้จากการค้นคืนสารสนเทศมีความถูกต้องเหมาะสมหรือไม่ ผู้วิจัยจะทำการทดลองด้วยคำหลักในการค้นคืนที่แตกต่างกัน 3 คำต่อหนึ่งชุดการทดสอบและจะทำการหาค่าเฉลี่ยของจำนวนผลลัพธ์ในการค้นคืนมาแสดง ซึ่งคำหลักที่ผู้วิจัยใช้ในการทดสอบมีสามคำได้แก่ คอมพิวเตอร์ ออนโทโลยีและสืบค้น โดยผู้วิจัยสรุปผลการทดสอบและรวบรวมผลการทดสอบซึ่งได้แสดงผลดังตารางที่ 4.1 4.2 และ 4.3

ตารางที่ 4.1 แสดงผลการทดสอบระบบโดยใช้คำหลักคือ คอมพิวเตอร์

ครั้งที่	จำนวนไฟล์	ผลลัพธ์ทั้งหมด		ผลลัพธ์ที่เกี่ยวข้อง		จำนวนผลลัพธ์ในระบบ
		Windows7	ระบบที่พัฒนา	Windows7	ระบบที่พัฒนา	
1	10	3	1	1	1	2
2	20	9	4	3	3	4
3	30	16	8	5	5	5
4	40	22	12	6	7	8
5	50	30	18	11	11	11

ตารางที่ 4.2 แสดงผลการทดสอบระบบโดยใช้คำหลักคือ ออนโทโลยี

ครั้งที่	จำนวนไฟล์	ผลลัพธ์ทั้งหมด		ผลลัพธ์ที่เกี่ยวข้อง		จำนวนผลลัพธ์ในระบบ
		Windows7	ระบบที่พัฒนา	Windows7	ระบบที่พัฒนา	
1	10	3	2	2	2	2
2	20	5	2	2	2	2
3	30	10	2	3	2	3
4	40	10	3	4	3	4
5	50	12	5	4	5	5

ตารางที่ 4.3 แสดงผลการทดสอบระบบโดยใช้คำหลักคือ สืบค้น

ครั้งที่	จำนวนไฟล์	จำนวนผลลัพธ์ที่ได้ทั้งหมด		จำนวนผลลัพธ์ที่ถูกต้อง		จำนวนผลลัพธ์ในระบบ
		Windows7	ระบบที่พัฒนา	Windows7	ระบบที่พัฒนา	
1	10	3	2	2	2	2
2	20	5	2	2	2	2
3	30	10	3	3	2	5
4	40	11	3	3	2	6
5	50	12	5	4	4	6

จากตารางที่ 4.1 4.2 และ 4.3 แสดงผลการทดสอบระบบค้นคืนสารสนเทศที่พัฒนาขึ้นโดยผู้วิจัยเปรียบเทียบกับระบบค้นคืนเอกสารในระบบปฏิบัติการ Windows 7 ในตารางแสดงให้เห็นว่าระบบที่พัฒนาขึ้น มีการแสดงจำนวนผลลัพธ์ของการค้นคืนที่น้อยกว่าการค้นคืนของระบบปฏิบัติการ Windows 7 ในทุกๆครั้งของการทดสอบ โดยเมื่อมีการเพิ่มจำนวนเอกสารเข้าระบบมากขึ้น ผลลัพธ์ในการค้นคืนของระบบปฏิบัติการ Windows 7 จะมีการเพิ่มจำนวนขึ้นเป็นอย่างมาก ทำให้เกิดปัญหาสำหรับผู้ใช้งานในการคัดกรองข้อมูลเพื่อนำมาใช้งาน ในส่วนของระบบที่พัฒนาขึ้นนั้น เมื่อทำการเพิ่มจำนวนเอกสารเข้าสู่ระบบจะมีการเพิ่มขึ้นของผลลัพธ์เช่นเดียวกัน แต่การเพิ่มขึ้นของผลลัพธ์นั้นมีจำนวนไม่สูงเท่ากับการค้นคืนด้วยระบบปฏิบัติการ Windows 7 ทำให้ผู้ใช้งานสามารถลดเวลาในการคัดกรองข้อมูลที่ได้เพื่อนำไปใช้งานรวดเร็วขึ้น

ในส่วนของการทดสอบประสิทธิภาพด้านความถูกต้องแม่นยำนั้นผู้วิจัยได้ทำการรวบรวมข้อมูลของการทดสอบ เพื่อนำไปหาค่าความแม่นยำ (Precision) และค่าการเรียกคืน (Recall) โดยมีสมการในการคำนวณออกมาดังนี้

Precision = Number of relevant documents retrieve / Total number of documents retrieve (4-1)

จากสมการที่ 4-1 คำนวณค่าความแม่นยำได้จากจำนวนผลลัพธ์ที่เกี่ยวข้องและถูกค้นคืนทั้งหมดหารด้วยจำนวนเอกสารทั้งหมดที่เป็นผลลัพธ์ของการค้นคืน

Recall = Number of relevant documents retrieve / Total number of relevant documents (4-2)

จากสมการที่ 4-2 คำนวณค่าการเรียกคืนได้จากจำนวนผลลัพธ์ที่เกี่ยวข้องและถูกค้นคืนทั้งหมดหารด้วยจำนวนเอกสารทั้งหมดที่เกี่ยวข้อง

เมื่อนำผลการทดลองจากตารางที่ มาคำนวณหาค่าตามสมการที่ 4-1 และสมการที่ 4-2 ผลที่ได้ของการค้นคืนสารสนเทศเมื่อเปรียบเทียบระหว่างการค้นคืนสารสนเทศด้วยระบบที่พัฒนาขึ้นกับการค้นคืนสารสนเทศด้วยระบบของระบบปฏิบัติการ Windows 7 ได้ผลดังแสดงในตารางที่ 4.4 4.5 และ 4.6

ตารางที่ 4.4 แสดงผลการคำนวณค่าความแม่นยำและค่าการเรียกคืนของการค้นคืนสารสนเทศโดยใช้คำหลักคือ คอมพิวเตอร์

การทดสอบครั้งที่	ค่าความแม่นยำ (Precision)		ค่าการเรียกคืน (Recall)	
	Windows 7	ระบบที่พัฒนาขึ้น	Windows 7	ระบบที่พัฒนาขึ้น
1	0.33	1	0.5	0.5
2	0.33	0.75	0.75	0.75

3	0.31	0.63	1	1
4	0.27	0.32	0.75	0.88
5	0.36	0.61	1	1
ค่าเฉลี่ย	0.32	0.66	0.8	0.83

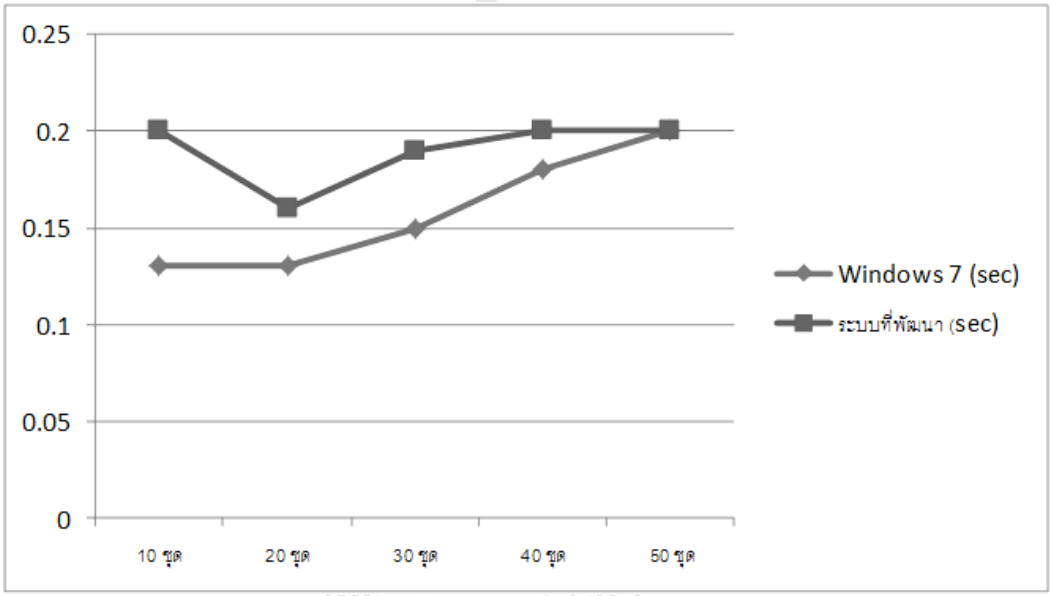
ตารางที่ 4.5 แสดงผลการคำนวณค่าความแม่นยำและค่าการเรียกคืนของการค้นคืนสารสนเทศโดยใช้คำหลักคือ ออนโทโลยี

การทดสอบครั้งที่	ค่าความแม่นยำ (Precision)		ค่าการเรียกคืน (Recall)	
	Windows 7	ระบบที่พัฒนาขึ้น	Windows 7	ระบบที่พัฒนาขึ้น
1	0.66	1	1	1
2	0.4	1	1	1
3	0.3	1	1	0.66
4	0.4	1	1	0.75
5	0.33	1	0.8	1
ค่าเฉลี่ย	0.42	1	0.96	0.88

ตารางที่ 4.6 แสดงผลการคำนวณค่าความแม่นยำและค่าการเรียกคืนของการค้นคืนสารสนเทศโดยใช้คำหลักคือ สืบค้น

การทดสอบครั้งที่	ค่าความแม่นยำ (Precision)		ค่าการเรียกคืน (Recall)	
	Windows 7	ระบบที่พัฒนาขึ้น	Windows7	ระบบที่พัฒนาขึ้น
1	0.66	1	1	1
2	0.4	1	1	1
3	0.3	0.66	0.6	0.4
4	0.27	0.66	0.5	0.33
5	0.42	0.8	0.66	0.66
ค่าเฉลี่ยรวม	0.41	0.82	0.75	0.68

ในส่วนสุดท้ายของการทดสอบประสิทธิภาพของระบบ ผู้วิจัยได้ทำการทดสอบประสิทธิภาพด้านเวลาของการค้นคืนสารสนเทศ โดยเป็นการเปรียบเทียบเวลาระหว่างระบบที่พัฒนาขึ้นกับระบบค้นคืนค้นคืนสารสนเทศด้วยระบบของระบบปฏิบัติการ Windows 7 ซึ่งผู้วิจัยใช้คำหลักในการค้นคืนคือคำว่าคอมพิวเตอร์ โดยแสดงเป็นแผนภูมิเปรียบเทียบด้านเวลาดังแสดงในรูปที่ 4.11



รูปที่ 4.11 แผนภูมิเปรียบเทียบเวลาในการแสดงผลลัพธ์ของการค้นคืนด้วยคำว่า คอมพิวเตอร์

จากรูปที่ 4.11 แสดงแผนภูมิเปรียบเทียบเวลาในการแสดงผลลัพธ์ของการค้นคืนสารสนเทศระหว่างระบบที่ผู้วิจัยพัฒนาขึ้น และระบบค้นคืนสารสนสนเทศของระบบปฏิบัติการ Windows7 โดยผู้วิจัยได้ใช้คำหลักคือคำว่า คอมพิวเตอร์ ซึ่งจากการทดสอบประสิทธิภาพด้านเวลาผลที่ได้พบว่าระบบที่ผู้วิจัยพัฒนาขึ้น เมื่อมีการเพิ่มขึ้นของจำนวนเอกสาร เวลาในการค้นคืนและแสดงผลลัพธ์ของระบบมีการเปลี่ยนแปลงเพียงเล็กน้อย แสดงว่าจำนวนเอกสารมีผลต่อการใช้เวลาในการแสดงผลลัพธ์ของระบบที่ผู้วิจัยพัฒนาขึ้นเพียงเล็กน้อย เมื่อเปรียบเทียบกับระบบค้นคืนสารสนเทศของระบบปฏิบัติการ Windows7 จะพบว่าเมื่อมีการเพิ่มขึ้นของจำนวนเอกสาร เวลาที่ใช้

ในการค้นคืนผลลัพธ์มีลักษณะการเพิ่มขึ้นเรื่อยๆ แสดงว่าเมื่อมีจำนวนเอกสารเพิ่มมากขึ้น ระบบปฏิบัติการ Windows7 มีแนวโน้มที่จะใช้เวลาในการแสดงผลลัพธ์มากขึ้นไปด้วย

4.4 การอภิปรายผล

ในการพัฒนาและทดสอบการใช้งานและประสิทธิภาพของระบบที่ได้พัฒนาขึ้น แบ่งเป็นสองส่วนคือ ส่วนของผู้ใช้งานทั่วไปและส่วนของผู้ดูแลระบบ ในส่วนการทดสอบของผู้ใช้งานทั่วไปนั้นระบบสามารถทำงานได้อย่างสมบูรณ์ มีหน้าจอในส่วนการแสดงผลลัพธ์ที่เข้าใจง่ายอีกทั้งยังมีส่วนที่ช่วยในการสะกดคำและส่วนที่แนะนำคำหลักในการค้นคืน ที่ใกล้เคียงกับการค้นคืนในแต่ละครั้ง ทำให้สะดวกในการค้นคืนมากขึ้น ทำให้ผู้ใช้งานสามารถใช้งานในการค้นคืนสารสนเทศได้สะดวกรวดเร็วในการคัดกรองข้อมูลเพื่อนำไปใช้งานได้รวดเร็วยิ่งขึ้น ในส่วนของการทดสอบการใช้งานของผู้ดูแลระบบนั้น พบว่าระบบมีเมนูที่ช่วยในการปรับปรุงข้อมูลของระบบได้สะดวก ไม่ว่าจะเป็นการเพิ่มลดคำศัพท์ในระบบหรือการเพิ่มลดผู้ดูแลระบบ รวมถึงการแก้ไขคำศัพท์ในระบบปรับปรุงดัชนีในการค้นคืนสารสนเทศ ผู้ดูแลระบบสามารถจัดการกับระบบได้ง่ายและรวดเร็ว ในส่วนการทดสอบประสิทธิภาพของการค้นคืนสารสนเทศ โดยทำการเปรียบเทียบกับระบบค้นคืนสารสนเทศที่มาพร้อมกับระบบปฏิบัติการ Windows7 จากผลการทดสอบถูกนำมาคำนวณหาค่าความแม่นยำ (Precision) และค่าการเรียกคืน (Recall) จากการคำนวณที่ได้มา ระบบที่พัฒนาขึ้น มีค่าความแม่นยำที่สูงกว่าระบบค้นคืนสารสนเทศที่มาพร้อมกับระบบปฏิบัติการ Windows7 โดยมีค่าความแม่นยำเฉลี่ยอยู่ระหว่าง 0.66 – 1.00 ซึ่งสามารถกล่าวได้ว่าระบบที่พัฒนาขึ้นสามารถค้นคืนสารสนเทศได้ดี ผลลัพธ์ที่ได้มีความน่าเชื่อถือสูง ในส่วนของการคำนวณหาค่าการเรียกคืนนั้น เมื่อคำนวณออกมาค่าการเรียกคืนของระบบที่พัฒนาขึ้นจะมีค่าเฉลี่ยอยู่ระหว่าง 0.68 – 0.83 ซึ่งถือว่าเป็นค่าการเรียกคืนที่สูง ผลลัพธ์ที่ได้มีความน่าเชื่อถือสูง สามารถนำไปใช้งานได้อย่างดี ส่วนสุดท้ายคือการเปรียบเทียบด้านเวลาในการแสดงผลลัพธ์ พบว่าเวลาในการค้นคืนผลลัพธ์ของระบบที่พัฒนาขึ้น เมื่อเปรียบเทียบกับระบบค้นคืนสารสนเทศของระบบปฏิบัติการ Windows7 มีการใช้เวลาใกล้เคียงกัน แต่เมื่อมีการเพิ่มจำนวนเอกสารขึ้น ระบบค้นคืนสารสนเทศของระบบปฏิบัติการ Windows7 มีแนวโน้มจะใช้เวลาเพิ่มขึ้น แต่ในส่วนของระบบที่ผู้วิจัยพัฒนาขึ้นเวลาที่ใช้แสดงผลลัพธ์ยังคงไม่มีแนวโน้มที่จะเปลี่ยนแปลงตามจำนวนเอกสาร

หลังจากได้ทำการพัฒนาและทดสอบระบบซึ่งได้ผลการทดสอบจริงของระบบออกมา ในบทต่อไปผู้วิจัยจะได้อธิบายถึงบทสรุปผลการวิจัยและข้อเสนอแนะ โดยจะมีรายละเอียดที่สำคัญคือ สรุปผลการวิจัย ประโยชน์ของระบบ ข้อจำกัดของระบบและแนวทางในการพัฒนาต่อ

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

งานวิจัยนี้มุ่งศึกษาค้นคว้า ตลอดจนนำเสนอแนวทางต้นแบบและพัฒนาระบบที่เหมาะสม เพื่อช่วยลดต้นทุนให้กับองค์กรต่างๆ ในการพัฒนาระบบค้นคืนสารสนเทศด้านเทคโนโลยีสารสนเทศภาษาไทย โดยใช้ความสามารถของการพัฒนาโปรแกรมในรูปแบบของการใช้หลักการเชิงความความหมายมาเป็นแนวทางในการพัฒนาระบบเว็บแอพพลิเคชันให้ดีขึ้น เพราะฉะนั้นจุดประสงค์ของงานวิจัยนี้คือช่วยลดเวลาในการคัดกรองผลลัพธ์จากการค้นคืนสารสนเทศ ลดค่าใช้จ่ายภายในองค์กรและอำนวยความสะดวกรวดเร็วให้กับผู้ใช้งาน

ขั้นตอนการดำเนินงานวิจัยนี้จะแบ่งออกเป็นหลักๆ ได้ 3 ส่วนด้วยกันคือ

1. การศึกษาข้อมูลที่เกี่ยวข้องกับการตัดในคำภาษาไทยและการค้นคืนสารสนเทศโดยใช้หลักการเชิงความหมาย เพื่อหาข้อมูลที่จำเป็นสำหรับการพัฒนาและออกแบบระบบ 2. การพัฒนาและทดสอบระบบ ผู้วิจัยได้พัฒนาระบบค้นคืนสารสนเทศด้านเทคโนโลยีสารสนเทศภาษาไทยขึ้นซึ่งมีความครบถ้วนและสมบูรณ์ในส่วนของรายละเอียดต่างๆ ในการใช้งาน ซึ่งรายละเอียดต่างๆ ที่นำมาพัฒนานั้นได้มาจากการสรุปผลในการค้นคว้าข้อมูล และ 3. การทดสอบประสิทธิภาพของระบบโดยทำการเปรียบเทียบผลลัพธ์ของการค้นคืนสารสนเทศ ระหว่างระบบที่นิยมใช้งานอย่างแพร่หลายในปัจจุบันกับระบบที่ผู้วิจัยได้พัฒนาขึ้น

โดยในบทนี้ผู้วิจัยจะได้อธิบายถึงผลสรุปต่างๆ ที่ได้จากการทดสอบระบบที่พัฒนาขึ้น โดยแบ่งออกเป็นหัวข้อดังนี้ สรุปผลการวิจัย ประโยชน์ของระบบ ข้อจำกัดของระบบและแนวทางในการพัฒนาต่อ

5.1 สรุปผลการวิจัย

ในกระบวนการพัฒนาระบบค้นคืนสารสนเทศด้านเทคโนโลยีสารสนเทศภาษาไทยโดยใช้หลักการเชิงความหมาย ผู้วิจัยจำเป็นต้องค้นคว้าข้อมูลต่างๆ เพื่อให้ครอบคลุมกับการใช้งานจริง โดยจากการวิจัยและทดสอบประสิทธิภาพของระบบ ทำให้สรุปได้ว่าระบบที่ผู้วิจัยได้พัฒนาขึ้นนี้ช่วยเพิ่มประสิทธิภาพในการค้นคืนสารสนเทศให้กับผู้ใช้งาน ลดระยะเวลาในการค้นคืนสารสนเทศได้มากขึ้นผลลัพธ์ที่ได้มีความถูกต้องแม่นยำมากขึ้น โดยระบบที่พัฒนาขึ้นยังสามารถใช้งานได้ในทุกระบบปฏิบัติการเนื่องจากพัฒนาด้วยภาษาจาวา

ในส่วนของผู้ใช้งานสามารถที่จะทำการค้นคืนสารสนเทศได้ ผลลัพธ์ที่ระบบแสดงออกมา ผู้ใช้งานสามารถที่จะเปิดดูฉบับเต็มได้ ตรวจสอบไดเรคทอรีที่เก็บเอกสารได้หรือสามารถบันทึกเอกสารที่ต้องการเพื่อนำไปศึกษาต่อได้

ในส่วนของผู้ดูแลระบบจะมีหน้าที่ทำการสร้างดัชนีของการค้นคืนสารสนเทศ ปรับปรุงแก้ไขเอกสารให้มีความทันสมัย เพิ่ม หรือลบคำศัพท์ในระบบและจัดการเพิ่มหรือลดผู้ดูแลระบบ

ในส่วนของผู้ดูแลระบบสูงสุดจะมีหน้าที่คล้ายกับผู้ดูแลระบบทุกประการ แต่ส่วนที่ผู้ดูแลระบบสูงสุดมีหน้าที่มากกว่าผู้ดูแลระบบคือ การเพิ่มหรือลดผู้ดูแลระบบ และกำหนดสิทธิ์ให้ผู้ดูแลระบบเป็นผู้ดูแลระบบสูงสุด

5.2 ประโยชน์ของระบบ

1. ช่วยอำนวยความสะดวกในการค้นคืนสารสนเทศที่เป็นภาษาไทยโดยเฉพาะ
2. มีการจัดเก็บคำศัพท์ด้านเทคโนโลยีสารสนเทศภาษาไทยไว้ให้พร้อมใช้งาน อีกทั้งยังสามารถทำการเพิ่มคำศัพท์ใหม่เข้าสู่ระบบได้ ทำให้ระบบมีความทันสมัยอยู่ตลอดเวลา
3. สามารถนำระบบที่พัฒนาขึ้นไปใช้ได้ในทุกองค์กร เพียงนำไปติดตั้งในระบบเว็บเซิร์ฟเวอร์ก็สามารถใช้งานได้ทันที
4. สามารถนำระบบนี้ไปประยุกต์ใช้กับงานด้านอื่นๆ ที่นอกเหนือจากด้านเทคโนโลยีสารสนเทศได้ โดยการเพิ่มคำศัพท์เฉพาะด้านเข้าสู่ระบบ เช่น ระบบค้นคืนสารสนเทศด้านชีววิทยาเป็นต้น

5.3 ข้อจำกัดของระบบ

1. คำศัพท์ต้องมีการปรับปรุงให้ทันสมัยตลอดเวลา เนื่องจากเป็นการพัฒนาในภาษาไทย อาจเกิดปัญหาในการนำไปใช้ในด้านอื่นที่นอกเหนือจากด้านเทคโนโลยีสารสนเทศ เนื่องจากคำในภาษาไทยอาจมีความหมายแตกต่างกันตามลักษณะการใช้งาน
2. เนื่องจากคำศัพท์ในระบบมีคำที่เขียนเลียนเสียงภาษาอังกฤษเป็นจำนวนมาก อาจเกิดปัญหาในการใช้งานของผู้ใช้งาน เนื่องจากความเข้าใจในการเขียนคำเลียนเสียงภาษาอังกฤษอาจแตกต่างกัน
3. ในบางครั้งในการค้นคืนสารสนเทศด้วยภาษาไทยนั้นอาจได้ผลลัพธ์ที่มีความถูกต้องไม่ดีเท่าที่ควร เนื่องจากระบบพัฒนาขึ้นโดยอาศัยการตัดคำเข้ามาช่วยในการสร้างดัชนีสำหรับการค้นคืนสารสนเทศ ซึ่งคำหลักที่ผู้ใช้งานระบุเข้ามาเพื่อทำการค้นคืนสารสนเทศอาจมีหลายความหมาย

และมีการใช้งานได้หลากหลายรูปแบบ ทำให้อาจเกิดข้อผิดพลาดในการสร้างดัชนีซึ่งมีผลต่อผลลัพธ์ในการค้นคืนสารสนเทศ

5.4 แนวทางในการพัฒนาต่อ

สำหรับแนวทางวิจัยที่จะพัฒนาต่อนั้นสามารถแบ่งออกเป็นได้หลายแนวทางดังนี้

1. พัฒนาในด้านของรูปแบบการแสดงผล เช่นอาจเพิ่มรูปภาพหรือสัญลักษณ์ในการแสดงผลลัพธ์ของการค้นคืนสารสนเทศ
2. เพิ่มเติมในส่วนของการค้นคืน เช่น ค้นหาด้วยชื่อผู้แต่ง ค้นหาจากชื่อรูปประกอบเป็นต้น
3. พัฒนาให้ผู้ดูแลระบบสามารถดูประวัติการแก้ไขคำศัพท์ในระบบได้
4. พัฒนาให้ผู้ดูแลระบบสามารถทำการสร้างดัชนีใหม่โดยการสร้างดัชนีเฉพาะเอกสารที่มีการเปลี่ยนแปลงเพื่อช่วยลดเวลาในการสร้างดัชนี

รายการอ้างอิง

Alexander Maedche and Steffen Staab. (2001). **Ontology Learning for the Semantic Web**. University of Karlsruhe.

Antoniou G. and Van Harmelen F. (2004). **A Semantic Web Primer**. Cambridge: MIT Press.

Brickley D. and Guha R.V. (2004). **Resource Description Framework (RDF) Schema Specification 1.0**. [Online]. Available: http://www.w3.org. [10 October 2010].

Daconta M., Obrst L. and Smith K. (2003). **The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management**. Indianapolis: Wiley.

David Martin, Mark Burstein, Jerry Hobbs, Ora Lassila, Drew McDermott, Sheila McIlraith, Srini Narayanan, Massimo Paolucci, Bijan Parsia, Terry Payne, Evren Sirin, Naveen Srinivasan and Katia Sycara. (2004). **OWL Web Ontology Language for Services (OWL-S)**. [Online]. Available: http://www.w3.org/Submission/OWL-S. [11 October 2010].

Junaidah Mohamed kassim and Mahathir Rahmany. (2009). **Introduction to Semantic Search Engine**. Electrical Engineering and Informatics, 2009. ICEEI '09. International Conference on 5 August, 2009.

Knublauch H., Ray W. Fergerson, Natalya F. Noy and Mark A. Musen. (2004). **The Protégé OWL Plug-in: An Open Development Environment for Semantic Web Applications**. Proceedings of the Semantic Web-ISWC 2004.

Mark W. (1995). **The computer for the 21st century**. Human-computer interaction: Toward the year 2000.

McBride B. (2002). **Jena: a semantic Web toolkit**. Internet Computing, IEEE 6th.

Reza Hemayati, Weiyi Meng and Clement Yu. (2007). **Semantic-Based Grouping of Search Engine Results Using WordNet**. APWeb/WAIM 2007.

Singh M. and Huhns M. (2005) **Service-oriented Computing - Semantic, Processes, Agents**. New York: John Wiley & Sons, Ltd.

Tim Berners-Lee, James Hendler and Ora Lassila. (2001). **The semantic Web**. Scientific American Magazine May 2001.

Tim Finin, Li Ding, Anupam Joshi, Yun Peng, R. Scott Cost, Joel Sachs, Rong Pan, Pavan Reddivari and Vishal Doshi (2004). **Swoogle: A Semantic Web Search and Metadata Engine**.

Varelas G., Voutsakis E., Raftopoulou P., Petrakis E. and Milios E. **Semantic similarity methods in WordNet and their application to information retrieval on the web**. Proceedings of the 7th annual ACM international workshop on Web information and data management (WIDM'05).

Yi Jin, Zhuying Lin and Hongwei Lin. (2008). **The Research of Search Engine Based on Semantic Web**. International Symposium on Intelligent Information Technology Application Workshops 2008.

ดวงแก้ว สวามิภักดิ์, **การสร้างซอฟต์แวร์วิเคราะห์ไวยากรณ์ไทยภายใต้ระบบยูนิกซ์**. มหาวิทยาลัยธรรมศาสตร์.2533.

ธนกร หวังพิพัฒน์วงศ์, อานนท์ ไกรเสวกวิสัย และสราวุธิ ราษฎร์นิยม. (2009) **ระบบค้นหารูปภาพโดยใช้หลักการเว็บเชิงความหมาย**. มหาวิทยาลัยกรุงเทพ.

ไพศาล เจริญพรสวัสดิ์. **Feature-based Thai word segmentation**. จุฬาลงกรณ์มหาวิทยาลัย. 2542.

วิรัศ ศรเลิศล้ำวณิช (2003). **การตัดคำไทยในระบบแปลภาษา**. การแปลภาษาด้วย คอมพิวเตอร์: โครงการพัฒนาระบบเครื่องแปลภาษาสำหรับภาษาในเอเชีย. กรุงเทพฯ : ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) 2003.

วิศิษฏ์ วรรณภูมิ และศิฬาณี นุชิตประสิทธิ์ชัย. (2009). **การสืบค้นข้อมูลการให้บริการเว็บเซอร์วิสเชิงความหมาย**. The 5th National Conference on Computing and Information Technology 2009.

สิทธิศักดิ์ บุญมาก (2006). **Search Engine คืออะไร**. [ออนไลน์].ได้จาก: http://www.pirun.ku.ac.th. [15 September 2010].

ภาคผนวก ก

รูปแบบหลักของการพัฒนาระบบโดยใช้ Lucene 3.0

รูปแบบการสร้างดัชนีของ Lucene3.0 โดยภาษาจาวา

ในการวิจัยครั้งนี้ได้มีการพัฒนาระบบโดยการเขียนโปรแกรมภาษาจาวา ซึ่งแบ่งเป็นส่วนหลักๆ ที่สำคัญคือ ส่วนของการสร้างดัชนีในการค้นคืนสารสนเทศและการค้นคืนสารสนเทศจากคำหลักที่ผู้ใช้งานระบุ

การสร้างดัชนีในการค้นคืนสารสนเทศของ Lucene3.0 โดยภาษาจาวานั้น ในขั้นแรกจะต้องมีการสร้างดัชนีจากเอกสารที่ต้องการนำมาใช้ในระบบ โดยขั้นแรกจะต้องมีการนำสารสนเทศที่ต้องการสร้างดัชนีไปเก็บไว้ในโฟล์เดอร์เดียวกันทั้งหมดก่อน หลังจากนั้นจะเป็นการเขียนโปรแกรมเพื่อทำการสร้างดัชนี โดยมีรายละเอียดของ Source Code ดังนี้

```
import java.io.File;
import java.io.FileFilter;
import java.io.FileReader;
import java.io.IOException;
import java.io.Reader;
import java.util.ArrayList;
import org.apache.lucene.analysis.standard.StandardAnalyzer;
import org.apache.lucene.analysis.th.ThaiAnalyzer;
import org.apache.lucene.document.Document;
import org.apache.lucene.document.Field;
import org.apache.lucene.index.IndexWriter;
import org.apache.lucene.store.Directory;
import org.apache.lucene.store.FSDirectory;
import org.apache.lucene.util.Version;

public class Indexer {
        public static void main(String[] args) throws Exception{
                String  indexdir =
"C:/Users/Martin/workspace2/WebSearch/WebContent/indexdir";
                String datadir = "C:/Users/Martin/workspace2/WebSearch/WebContent/datadir";
                long start = System.currentTimeMillis();
                Indexer indexer = new Indexer(indexdir);
                int numIndexed;
                try{
                        numIndexed = indexer.index(datadir, new TextFilesFilter());
                }finally{
                        indexer.close();
                        }
                long end = System.currentTimeMillis();
        System.out.println("Indexing " + numIndexed + " files "+ (end-start)+ " milliseconds");
        }
        private IndexWriter writer;
        public Indexer(String indexDir) throws Exception {
        Directory dir = FSDirectory.open(new File(indexDir));
        Writer = new IndexWriter
        (dir, new ThaiAnalyzer(),true,IndexWriter.MaxFieldLength.UNLIMITED);
          }
        public void close() throws Exception{
                writer.close();
          }

        public int index(String dataDir, FileFilter filter)throws Exception{
                File[] files = new File(dataDir).listFiles();
```

```
                for(File f: files){
                        if(!f.isDirectory()&&
                        !f.isHidden()&&
                        f.exists()&&
                        f.canRead()&&
                        (filter == null||filter.accept(f))){
                        indexFile(f);
                        }
                }
                return writer.numDocs();
        }

static class TextFilesFilter implements FileFilter{
        public boolean accept(File path){
                return path.getName().toLowerCase() != null;
                                        }
                                }
protected Document getDocument(File f) throws Exception{
        Document doc = new Document();
        doc.add(new Field("contents", new FileReader(f)));
        doc.add(new Field("filename",f.getName(),
                Field.Store.YES, Field.Index.NOT_ANALYZED));
        doc.add(new Field("fullpath",f.getCanonicalPath(),
                Field.Store.YES, Field.Index.NOT_ANALYZED ));
        return doc;
        }
private void indexFile(File f) throws Exception{
        System.err.println("Indexing "+f.getCanonicalPath());
        Document doc =getDocument(f);
        writer.addDocument(doc);
            }
}
```

ในส่วนของการสร้างดัชนีสำหรับการค้นคืนสารสนเทศนั้น รายละเอียดที่สำคัญส่วนใหญ่จะเป็นคำสั่งที่ใช้เรียกดูเอกสารที่ต้องการจัดทำดัชนีและคำสั่งที่ใช้สำหรับบันทึกไฟล์ดัชนีที่สร้างเสร็จแล้ว โดยคำสั่งส่วนใหญ่เป็นการเรียกใช้คลาสของ Lucene 3.0 อีกส่วนที่สำคัญคือส่วนการกำหนดชนิดของเอกสารที่ต้องการให้ระบบทำการสร้างดัชนี ซึ่งในส่วนนี้ผู้วิจัยได้กำหนดชนิดของเอกสารโดยเลือกตามความนิยมของเอกสารด้านเทคโนโลยีสารสนเทศ ดังนี้ PDF, DOC, HTML และ TXT

ในส่วนถัดไปจะเป็นส่วนที่สำคัญอีกส่วนหนึ่งสำหรับระบบค้นคืนสารสนเทศ ซึ่งก็คือส่วนค้นคืนสารสนเทศตามคำหลักที่ผู้ใช้งานระบุ โดยผู้วิจัยได้พัฒนาขึ้นดังตัวอย่างของ Source Code ด้านล่างนี้

```
import java.awt.List;
import java.io.File;
import java.io.IOException;
import java.util.ArrayList;
import javax.servlet.http.HttpServlet;
import javax.servlet.http.HttpServletRequest;
import javax.servlet.http.HttpServletResponse;
```

```
import javax.servlet.http.HttpSession;
import org.apache.lucene.analysis.standard.StandardAnalyzer;
import org.apache.lucene.analysis.th.ThaiAnalyzer;
import org.apache.lucene.document.Document;
import org.apache.lucene.index.IndexReader;
import org.apache.lucene.index.Term;
import org.apache.lucene.queryParser.MultiFieldQueryParser;
import org.apache.lucene.queryParser.ParseException;
import org.apache.lucene.queryParser.QueryParser;
import org.apache.lucene.search.IndexSearcher;
import org.apache.lucene.search.Query;
import org.apache.lucene.search.ScoreDoc;
import org.apache.lucene.search.TopDocs;
import org.apache.lucene.search.WildcardQuery;
import org.apache.lucene.search.highlight.Encoder;
import org.apache.lucene.search.highlight.Formatter;
import org.apache.lucene.search.highlight.Highlighter;
import org.apache.lucene.search.highlight.InvalidTokenOffsetsException;
import org.apache.lucene.search.highlight.QueryScorer;
import org.apache.lucene.search.highlight.SimpleFragmenter;
import org.apache.lucene.search.highlight.SimpleHTMLEncoder;
import org.apache.lucene.search.highlight.SimpleHTMLFormatter;
import org.apache.lucene.search.spell.PlainTextDictionary;
import org.apache.lucene.search.spell.SpellChecker;
import org.apache.lucene.store.Directory;
import org.apache.lucene.store.FSDirectory;
import org.apache.lucene.util.Version;
import edu.smu.tspell.wordnet.NounSynset;
import edu.smu.tspell.wordnet.Synset;
import edu.smu.tspell.wordnet.SynsetType;
import edu.smu.tspell.wordnet.WordNetDatabase;
import edu.smu.tspell.wordnet.NounSynset;

public class Search {
        public String found;
        public ArrayList search(String q)throws IOException{
                System.out.println("key word in search = "+q);
                String result=null ;
                String indexDir
="C:/Users/Martin/workspace2/WebSearch/WebContent/indexdir";
                Directory dir = FSDirectory.open(new File(indexDir));
                IndexReader indexReader = IndexReader.open(dir);
                IndexSearcher is = new IndexSearcher(FSDirectory.open(new File(indexDir)));
                String [] field = {"filename","contents"};
            MultiFieldQueryParser parser = new MultiFieldQueryParser(field,new
ThaiAnalyzer());
                parser.setDefaultOperator(MultiFieldQueryParser.OR_OPERATOR);
                Query query = null;
                StringBuffer buffer = new StringBuffer();
                ArrayList array = new ArrayList();
                        try {
                                query = parser.parse(q);
                        long start = System.currentTimeMillis();
                        TopDocs hits = is.search(query,44);
                        long end = System.currentTimeMillis();
                        int i = 0;
                        for (ScoreDoc scoreDoc : hits.scoreDocs){
          Document doc = is.doc(scoreDoc.doc);
           Formatter formatter = new SimpleHTMLFormatter("<font
```

```
color=red><b>","</b></font>");
            QueryScorer scorer = new QueryScorer(query,"contents");
            Encoder encoder = new SimpleHTMLEncoder();
            Highlighter highlighter = new Highlighter(formatter,encoder,scorer);
            highlighter.setTextFragmenter(new SimpleFragmenter(200));
            result = highlighter.getBestFragment(new ThaiAnalyzer(), "filename",
doc.get("contents"));
            System.out.println( result);
            buffer.append("<font color=blue size=3 >");
            buffer.append("<b><form name=\"download"+i+"\" method=\"GET\"
action=\"Download\" target=\"_blank \"  >");
            buffer.append("<input type='hidden' name='filename'
value='"+doc.get("filename")+"'/>");
           buffer.append("<input type='hidden' name='path'
value='"+doc.get("filepath")+"'/></form>");
           buffer.append("<a href=\"#\"  onClick=\"javascript:document.download"+i+".submit();
           return false;\">");

buffer.append(doc.get("filename").substring(0,doc.get("filename").lastIndexOf('.')));
                        buffer.append("</a></b>");
                        buffer.append("</font>");
                        buffer.append("<br>");
                         buffer.append("<font  FACE=Arial, Helvetica,  sans-serif color=black>");
                        buffer.append(result);
                        buffer.append("</font>");
                        System.out.println(doc.get("contents"));
                        buffer.append("<font color=green size=2>");
                         buffer.append("<br>");
                         buffer.append(doc.get("filepath"));
                         buffer.append("<br>");
                         buffer.append("</font>");
                         buffer.append("<br>");
                                 i++;
                                 if(i%5 == 0){
                                         System.out.println("in if");
                                         array.add(buffer.toString());
                                         buffer.delete(0, buffer.length());
                                         }
                                 }
                                 if(i%5 != 0){
                                         array.add(buffer.toString());
                                 }

                               found = "<br>Found "+ i + " document(s)(in "+(end-start)+" milliseconds
)
                                 that matched keyword '<font color=red><b> " +q+ "</font></b> '";
                                   }catch(Exception ex){
                                         System.out.println("error");
                                         found = "";
                                 }
                        is.close();
                        result = buffer.toString();
                        System.out.println(array.size());
                        if (array.size()== 0){
                                  array.add("");
                        }
                  return array;
                }
```

ในส่วนนี้จะเป็นการรับค่าคำหลักที่ผู้ใช้งานระบุเข้ามา ซึ่งผู้วิจัยได้กำหนดตัวแปรไว้เพื่อเก็บค่าคำหลัก หลังจากนั้นจะใช้คำสั่งเพื่อนำคำหลักไปค้นหาสารสนเทศจากไดเรคทอรีที่เก็บไฟล์ดัชนี ซึ่งได้กำหนดไว้ในขั้นตอนการสร้างดัชนี โดยมีการเขียนคำสั่งเพื่อรองรับในกรณีไม่พบเอกสารที่ใกล้เคียงไว้ด้วย ในตัวอย่าง Source Code ที่ยกตัวอย่างมาข้างบนนี้ ผู้วิจัยได้มีการใส่คำสั่งในการแสดงผลไว้ด้วย ซึ่งผู้วิจัยได้กำหนดรายละเอียดของเอกสารที่ระบบสามารถค้นคืนได้ที่จะนำมาแสดงในหน้าจอแสดงผล คือ ชื่อเอกสาร รายละเอียดที่เกี่ยวข้องกับคำหลักและไดเรคทอรีที่เก็บเอกสาร ซึ่งผู้ใช้งานสามารถคลิกที่ชื่อไฟล์เพื่อเปิดเอกสารต้นฉบับได้

อีกส่วนหนึ่งของ Source Code ที่ผู้วิจัยจะได้ยกเป็นตัวอย่างเพื่อความเข้าใจในระบบค้นคืนเอกสารซึ่งก็คือ ส่วนของการเรียกใช้คลาสสำหรับอ่านไฟล์ที่ผ่านการเข้ารหัสชนิดต่างๆ เพื่อนำมาสร้างดัชนี ในส่วนนี้ผู้วิจัยได้ใช้ส่วนเสริมที่ชื่อ Tika Indexer ซึ่งสามารถดาวน์โหลดนำมาใช้งานโดยไม่มีค่าใช้จ่ายและถูกพัฒนาขึ้นมาสำหรับใช้งานกับ Lucene3.0 ได้เป็นอย่างดี ตัวอย่างในการสร้างดัชนีจากเอกสารที่ผ่านการเข้ารหัสชนิดต่างๆ สามารถทำได้ดังนี้

```
import java.io.File;
import java.io.FileFilter;
import java.io.FileInputStream;
import java.io.IOException;
import java.io.InputStream;
import java.util.ArrayList;
import java.util.Collections;
import java.util.Date;
import java.util.HashSet;
import java.util.Iterator;
import java.util.List;
import java.util.Set;
import org.apache.lucene.document.Document;
import org.apache.lucene.document.Field;
import org.apache.tika.config.TikaConfig;
import org.apache.tika.metadata.Metadata;
import org.apache.tika.parser.AutoDetectParser;
import org.apache.tika.parser.ParseContext;
import org.apache.tika.parser.Parser;
import org.apache.tika.sax.BodyContentHandler;
import org.apache.tika.sax.WriteOutContentHandler;
import org.xml.sax.ContentHandler;
import org.apache.pdfbox.util.PDFStreamEngine;
public class TikaIndexer extends Indexer  {
        private boolean DEBUG = false;
        private int maxStringLength = -1;
        static Set<String> textualMetadataFields = new HashSet<String>();
        static {
                        textualMetadataFields.add(Metadata.TITLE);
                        textualMetadataFields.add(Metadata.AUTHOR);
                        textualMetadataFields.add(Metadata.COMMENTS);
                        textualMetadataFields.add(Metadata.KEYWORDS);
```

```
                    textualMetadataFields.add(Metadata.DESCRIPTION);
                    textualMetadataFields.add(Metadata.SUBJECT);
                    textualMetadataFields.add(Metadata.CONTENT_DISPOSITION);
                    }
            public static void main(String[] arg) throws Exception {
            TikaConfig config = TikaConfig.getDefaultConfig();
            List parsers = new ArrayList(config.getParsers().keySet());
            Collections.sort(parsers);
            Iterator it = parsers.iterator();
            System.out.println("Mime type parsers:");
            while (it.hasNext()) {
                    System.out.println(" " + it.next());
            }
            System.out.println();
            String indexDir =
"C:/Users/Martin/workspace2/WebSearch/WebContent/indexdir";
            String dataDir =
"C:/Users/Martin/workspace2/WebSearch/WebContent/datadir";
            long start = new Date().getTime();
            TikaIndexer indexer = new TikaIndexer(indexDir);
            int numIndexed = indexer.index(dataDir, new TextFilesFilter());
            indexer.close();
            long end = new Date().getTime();
            System.out.println("Indexing " + numIndexed + " files took "
                            + (end - start) + " milliseconds");
    }
    public TikaIndexer(String indexDir) throws Exception {
            super(indexDir);
    }
    protected Document getDocument(File f) throws Exception {
            Metadata metadata = new Metadata();
            metadata.set(Metadata.RESOURCE_NAME_KEY, f.getName());
            InputStream is = new FileInputStream(f);
            WriteOutContentHandler handler = new WriteOutContentHandler(
                            maxStringLength);
            Parser parser = new AutoDetectParser();
            ParseContext context = new ParseContext();
            context.set(Parser.class, parser);
            try {
                    parser.parse(is, handler, metadata, new ParseContext());
            }catch(Exception ex){
                    System.out.println(ex.getMessage());
            }
            finally {
                    is.close();
            }
            Document doc = new Document();
            doc.add(new Field("contents", handler.toString(), Field.Store.YES,
                            Field.Index.ANALYZED));
            if (DEBUG) {
                    System.out.println(" all text: " + handler.toString());
            }
            for (String name : metadata.names()) {
                    String value = metadata.get(name);
                    if (textualMetadataFields.contains(name)) {
                            doc.add(new Field("contents", value, Field.Store.YES,
                                            Field.Index.ANALYZED));
                    }
                    doc.add(new Field(name, value, Field.Store.YES, Field.Index.NO));
                    if (DEBUG) {
```

```
                                System.out.println(" " + name + ": " + value);
                        }
                }
                if (DEBUG) {
                        System.out.println();
                }
                doc.add(new Field("filepath", f.getCanonicalPath(), Field.Store.YES,
                                Field.Index.NOT_ANALYZED));
                doc.add(new Field("filename", f.getName(), Field.Store.YES,
                                Field.Index.ANALYZED));
                return doc;
        }
        public int getMaxStringLength() {
                return maxStringLength;
        }
        public void setMaxStringLength(int maxStringLength) {
                this.maxStringLength = maxStringLength;
        }
}
```

จาก Source Code ที่ผู้วิจัยยกมาเป็นตัวอย่างนั้น เป็นส่วนที่จำเป็นสำหรับการสร้างระบบค้นคืนสารสนเทศที่พัฒนาขึ้นโดยใช้ Lucene 3.0 ซึ่งมีตั้งแต่การสร้างดัชนี การค้นคืนสารสนเทศจากคำหลักและการอ่านไฟล์ในชนิดที่แตกต่างกันเพื่อนำไปสร้างดัชนีสำหรับการค้นคืน การนำไปใช้งานนั้นขึ้นอยู่กับผู้ที่นำไปพัฒนาต่อว่าต้องการให้ระบบทำงานอย่างไร โดยอาจเพิ่มรายละเอียดในส่วนอื่นๆ เข้ามาเพื่อให้ทำงานได้ตรงต่อการใช้งานมากขึ้น หรืออาจลดรายละเอียดลงเพื่อการใช้งานที่เรียบง่ายรวดเร็วมากยิ่งขึ้น

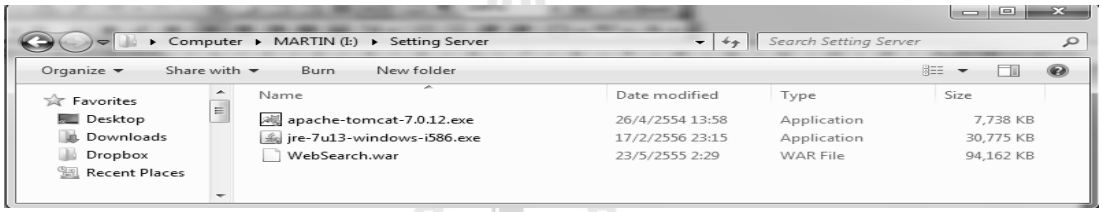
ภาคผนวก ข

วิธีการติดตั้งระบบ

ระบบค้นคืนสารสนเทศด้านเทคโนโลยีสารสนเทศภาษาไทย จำเป็นที่จะต้องมีเครื่องที่เก็บข้อมูลซึ่งเรียกว่าเครื่องกลางหรือเครื่อง Server ซึ่งจะเป็นเครื่องที่ใช้ในการติดตั้งระบบลงไป โดยในการติดตั้งระบบนั้นจะต้องมีโปรแกรมที่สำคัญคือ

- JRE (Java Runtime Environment) สำหรับประมวลผลภาษาจาวา
- Apache Tomcat ใช้สำหรับเชื่อมต่อระบบเครือข่ายในการใช้งาน
- WebSearch.war ซึ่งเป็นไฟล์สำหรับเริ่มต้นการใช้งานระบบ

ซึ่งโปรแกรมที่ต้องการนั้น ทางผู้วิจัยได้จัดเตรียมไว้ให้ผู้ที่สนใจสามารถดาวน์โหลดได้โดยเข้าไปที่ http://www.linux.sut.ac.th/download/Master_Thesis/Konjanapong/ โดยเมื่อเปิดมาจะพบไฟล์ที่มีทั้งหมดดังรูปที่ ค.1



รูปที่ ค.1 แสดงไฟล์ทั้งหมดที่จำเป็นในการติดตั้งระบบซึ่งมาพร้อมกับแผ่นซีดี

ในลำดับแรกให้ทำการติดตั้ง JRE ก่อน โดยดับเบิ้ลคลิกที่ไฟล์ jre-7u13-windows-i586.exe เมื่อเลือกแล้วจะได้หน้าจอดังรูปที่ ค.2

รูปที่ ค.2 แสดงหน้าจอโปรแกรมเมื่อเริ่มติดตั้ง

หลังจากนั้นให้กดปุ่ม Install เมื่อกดปุ่ม Install แล้วโปรแกรมจะแสดงหน้าจอดังรูปที่ ค.3

รูปที่ ค.3 แสดงหน้าจอคำเตือนและรายละเอียดของโปรแกรม

ให้กดปุ่ม Continue หลังจากนั้นโปรแกรมจะทำการติดตั้งดังรูปที่ ค.4

รูปที่ ค.4 แสดงหน้าจอโปรแกรมขณะทำการติดตั้ง

เมื่อโปรแกรมติดตั้งเสร็จจะขึ้นดังรูปที่ ก.5 แสดงว่าได้ทำการติดตั้ง JRE เสร็จเรียบร้อยแล้ว ให้ทำการกดปุ่ม close

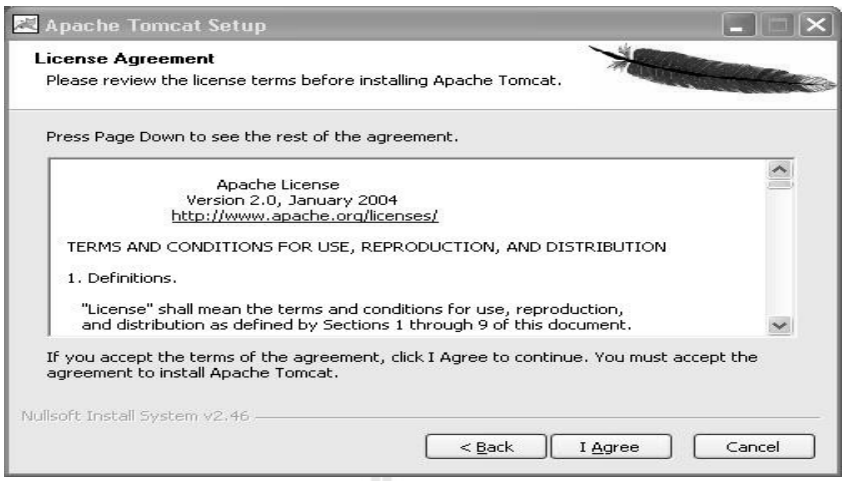
รูปที่ ค.5 แสดงหน้าจอโปรแกรมเมื่อติดตั้งสำเร็จ

หลังจากนั้นให้ทำการติดตั้ง Apache Tomcat โดยมีขั้นตอนดังนี้

ให้ทำการ Double Click ที่โปรแกรม apache-tomcat-7.0.12.exe ดังรูปเพื่อทำการแล้วจะได้หน้าจอดังรูปที่ ค.6

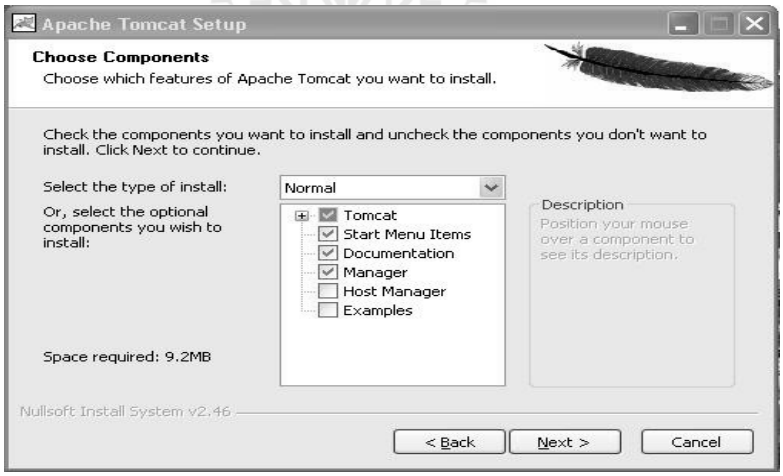
รูปที่ ค.6 แสดงหน้าจอโปรแกรมเมื่อเลือกติดตั้งโปรแกรม Apache Tomcat

หลังจากนั้นให้กดปุ่ม Next โปรแกรมจะแสดงหน้าจอดังรูปที่ ก.7



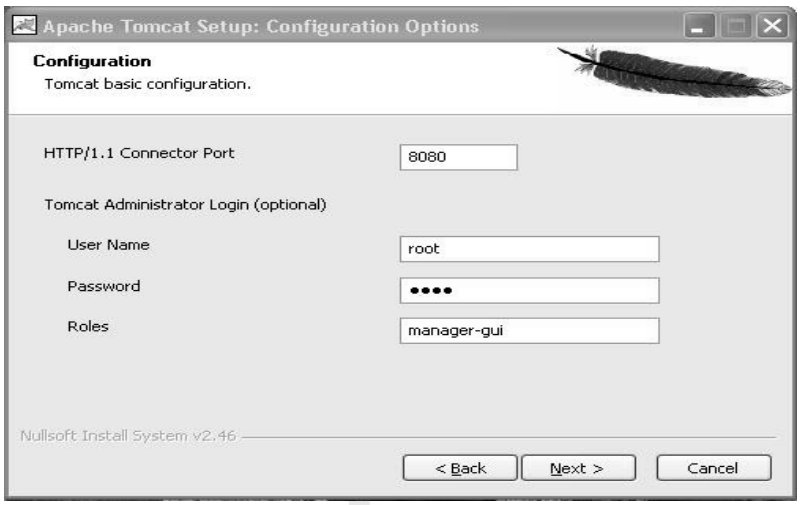
รูปที่ ค.7 แสดงหน้าจอรายละเอียดของโปรแกรม

หลังจากอ่านรายละเอียดเรียบร้อยแล้วให้กดปุ่ม I Agree โปรแกรมจะแสดงหน้าจอดังรูปที่ ค.8 ซึ่งเป็นหน้าจอแสดงรายละเอียดที่ระบบให้เลือกในการติดตั้ง



รูปที่ ค.8 แสดงหน้าจอรายละเอียดที่ต้องการติดตั้งของโปรแกรม

เมื่อพบหน้าจอดังรูปที่ ค.8 แล้วให้เลือกตามรูปหลังจากนั้นกดปุ่ม Next โปรแกรมจะแสดงหน้าจอดังรูปที่ ค.9



รูปที่ ค.9 แสดงหน้าจอในส่วนของการตั้งค่าโปรแกรม

เมื่อหน้าจอโปรแกรมแสดงดังรูปที่ ค.9 แล้ว ให้ผู้ใช้งานเลือกกำหนดค่าต่างๆ ดังนี้ ซึ่งผู้วิจัยได้กำหนดค่าเป็นตัวอย่างดังนี้

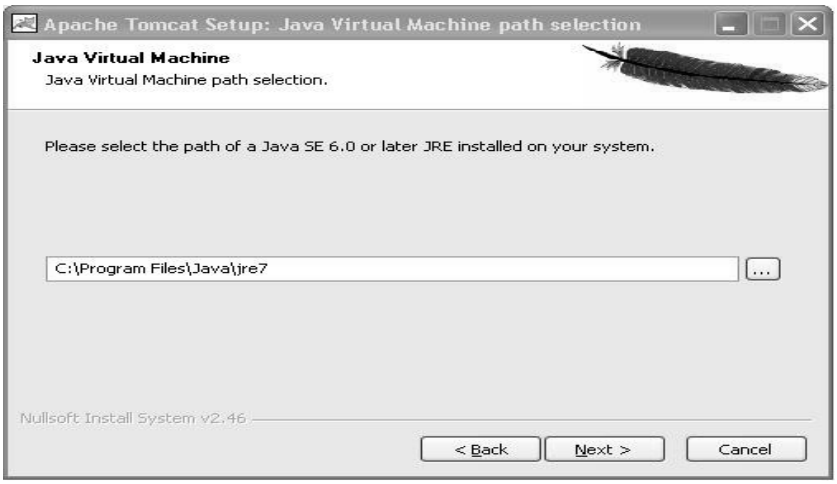
HTTP/1.1 Connector Port: 8080 (สามารถตั้งเป็นเลขอื่นได้)

User Name: root (สามารถตั้งเป็นชื่ออื่นได้)

Password: root (สามารถตั้งเป็นชื่ออื่นได้)

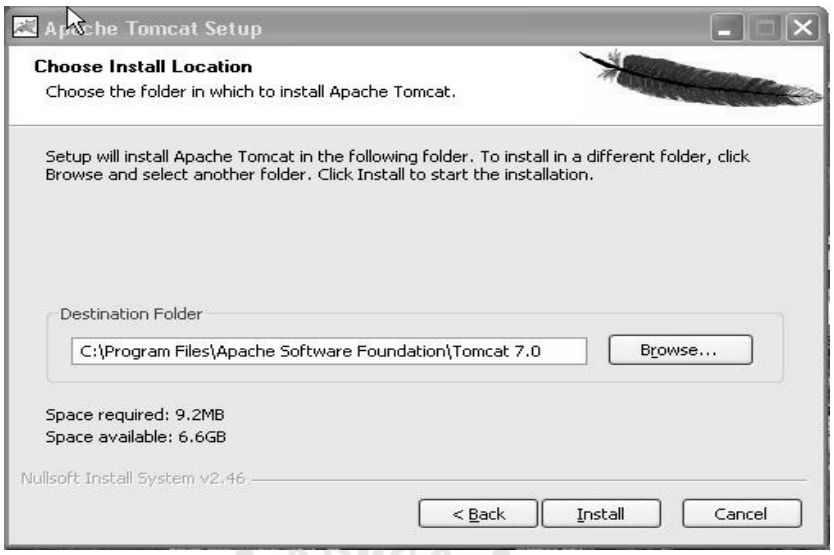
Roles: manager – gui (ตั้งเป็นค่านี้เท่านั้น)

หลังจากนั้นกดปุ่ม Next โปรแกรมจะแสดงหน้าจอดังรูปที่ ค.10



รูปที่ ค.10 แสดงหน้าจอโปรแกรมให้เลือกไดเรคทอรีที่เราติดตั้ง JRE เมื่อขั้นตอนแรก

ในส่วนนี้ให้เลือกไดเรคทอรีที่ได้ติดตั้ง JRE ไว้ในเครื่อง ในส่วนของผู้วิจัยได้ติดตั้งไว้ที่ C:\Program Files\java\jre7 เมื่อเลือกเสร็จแล้วกดปุ่ม Next โปรแกรมจะแสดงหน้าจอดังรูปที่ ค.11

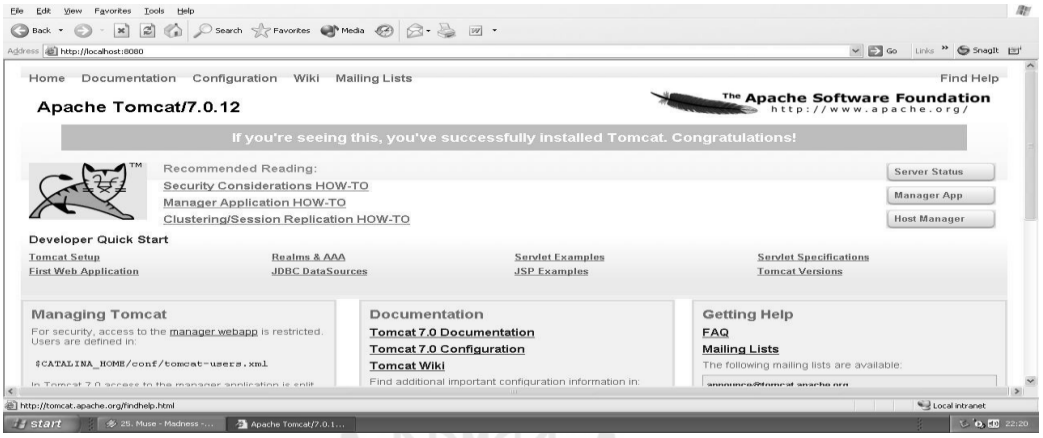


รูปที่ ค.11 แสดงหน้าจอโปรแกรมให้เลือก path ที่ต้องการติดตั้ง Apache Tomcat ในเครื่อง

ในส่วนนี้จะเป็นส่วนที่ต้องเลือกว่าจะติดตั้งโปรแกรม Apache Tomcat ลงไว้ที่ส่วนใดในเครื่อง โดยผู้วิจัยได้เลือกที่จะใช้ค่าพื้นฐานที่โปรแกรมกำหนดมาให้ดังรูป เมื่อเลือกแล้วให้กดปุ่ม Install เมื่อกดปุ่มเสร็จแล้วโปรแกรมจะแสดงหน้าจอดังรูปที่ ก.12

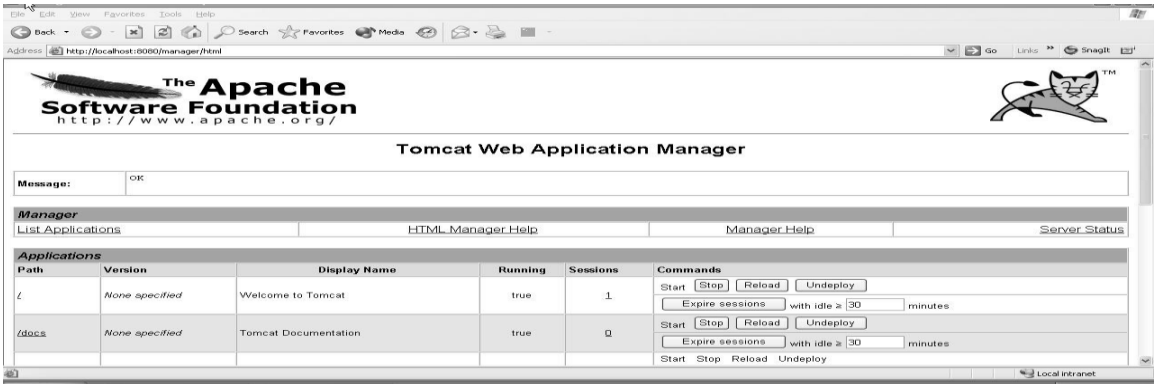
รูปที่ ค.12 แสดงหน้าจอโปรแกรมเมื่อติดตั้งเสร็จสมบูรณ์

เมื่อติดตั้งเสร็จสมบูรณ์แล้วต่อไปจะเป็นการทดสอบการติดตั้งว่าใช้งานได้หรือไม่ โดยเปิดโปรแกรม Web Browser ในเครื่องขึ้นมา ในส่วนนี้ผู้วิจัยเลือกใช้ Internet Explorer ในการทดสอบ ในช่องระบุ URL ให้ระบุดังนี้ http://localhost:8080 ถ้าระบบติดตั้งสมบูรณ์หน้าจอจะแสดงดังรูปที่ ค.13



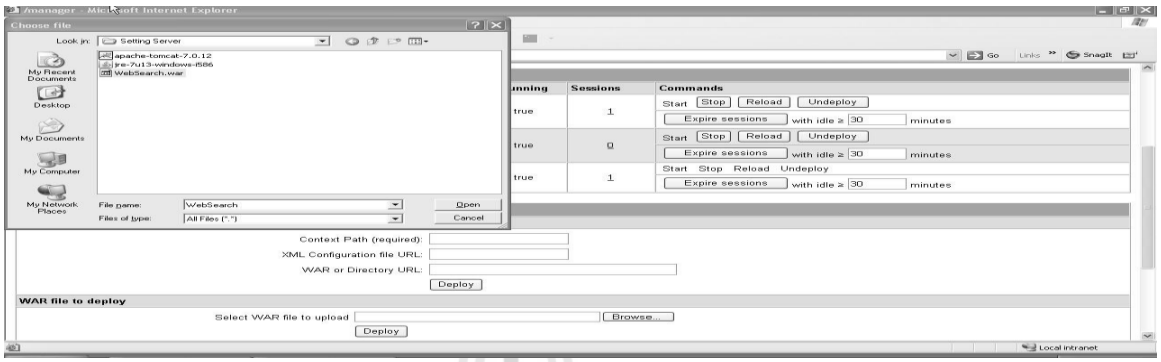
รูปที่ ค.13 แสดงหน้าจอการทำงานของโปรแกรม Apache Tomcat เมื่อติดตั้งสมบูรณ์

เมื่อโปรแกรมถูกติดตั้งสมบูรณ์ขั้นตอนต่อไปคือการนำไฟล์เริ่มต้นของระบบติดตั้งลงในเครื่องกลาง โดยเลือกที่ Manager App บนหน้าจอตามรูปที่ ค.13 หากเลือกแล้วระบบถามหาชื่อผู้ใช้งานและรหัสผ่าน ให้กรอกลงไปตามที่ได้เลือกไว้ตอนติดตั้งโปรแกรม Apache Tomcat ดังรูปที่ ก.9 เมื่อระบบทำการตรวจสอบชื่อผู้ใช้งานและรหัสผ่านว่าถูกต้องแล้วจะแสดงหน้าจอดังรูปที่ ค.14



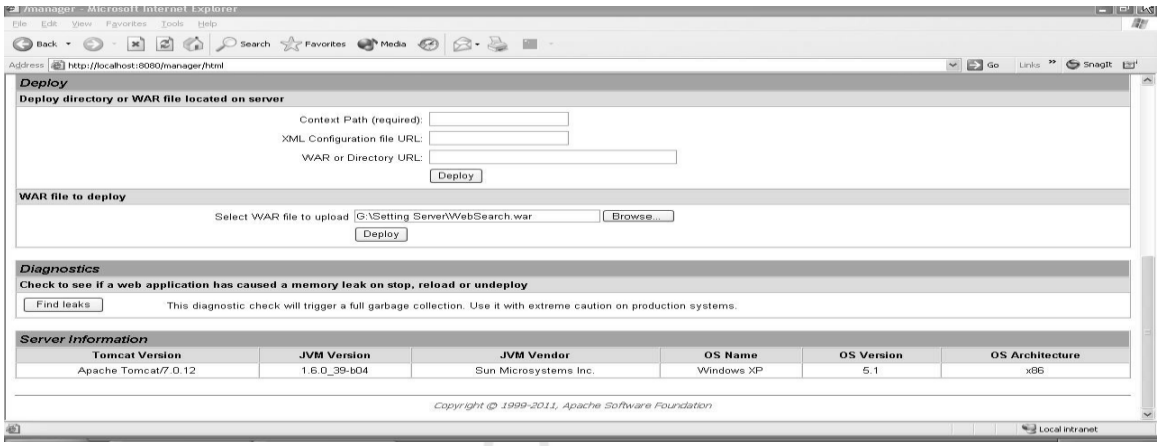
รูปที่ ค.14 แสดงหน้าจอเมื่อเข้าสู่ระบบสำเร็จ

หลังจากนั้นให้ทำการเลื่อนหน้าจอลงมาจนพบข้อความ“Select war file to upload” แล้วกด Browser หลังจากนั้นให้หาไฟล์ชื่อ WebSearch.war จากแผ่นซีดี แล้วกด open ดังรูปที่ ค.15



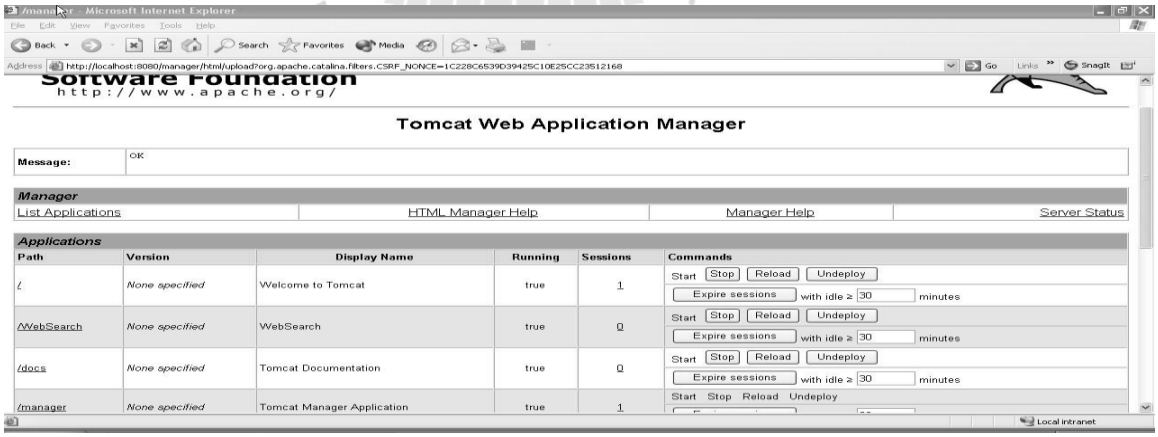
รูปที่ ค.15 แสดงหน้าจอขณะทำการเลือกไฟล์ WebSearch.war เพื่อทำการติดตั้งระบบ

หลังจากกด open แล้วโปรแกรมจะแสดงข้อความในช่อง ให้ตรวจสอบความถูกต้องอีกครั้งหนึ่งหลังจากตรวจสอบความถูกต้องแล้วให้กดปุ่ม Deploy ดังรูปที่ ค.16



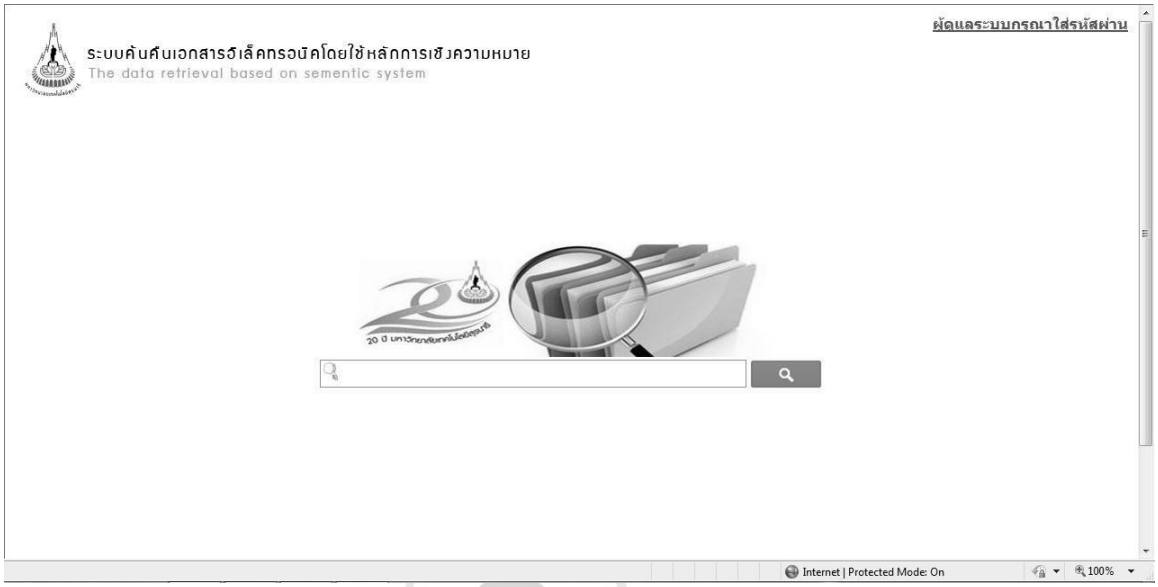
รูปที่ ค.16 แสดงหน้าจอหลังจากเลือกไฟล์ที่ต้องการติดตั้งจากซีดี

เมื่อกด Deploy แล้วโปรแกรมจะทำการประมวลผล เมื่อระบบทำการประมวลผลเสร็จสิ้นระบบจะแสดงหน้าจอดังรูปที่ ค.17



รูปที่ ค.17 แสดงหน้าจอเมื่อติดตั้งระบบสำเร็จมีรายละเอียดของไฟล์ WebSearch

หลังจากติดตั้งระบบสมบูรณ์แล้ว ให้ทำการทดสอบการใช้งานระบบโดยการเปิดโปรแกรม Web Browser แล้วทำการระบุ URL ดังนี้ http://localhost:8080/WebSearch หากระบบทำการติดตั้งสมบูรณ์ระบบจะแสดงหน้าจอดังรูปที่ ค.18



รูปที่ ค.18 แสดงหน้าจอระบบเมื่อติดตั้งสมบูรณ์พร้อมใช้งาน

ภาคผนวก ค

บทความวิชาการที่ได้รับการตีพิมพ์เผยแพร่ในระหว่างศึกษา

บทความวิชาการที่ได้รับการตีพิมพ์เผยแพร่ในระหว่างศึกษา

โกญจนพงษ์ ทองเพชร และ คะชา ชาญศิลป์ (2554) **แนวคิดในการค้นคืนข้อมูลเชิงความหมายด้านเทคโนโลยีสารสนเทศภาษาไทย**. การประชุมวิชาการระดับชาติมหาวิทยาลัยราชภัฏนครปฐม ครั้งที่ 3 (The 3rd NPRU National Conference 2011). หน้า 639 – 645.

75 ปี จากโรงเรียนสตรีศึกษัดครู สู่มหาวิทยาลัยเพื่อการพัฒนาท้องถิ่นอย่างยั่งยืน ครั้งที่ 3

Menu

ผู้สนับสนุน

มหาวิทยาลัยราชภัฏนครปฐม มหาวิทยาลัยราชภัฏเพชรบุรี มหาวิทยาลัยราชภัฏหมู่บ้านจอมบึง มหาวิทยาลัยราชภัฏกาญจนบุรี

เครือข่ายสหวิทยาการแห่งราชบัณฑิตยสถาน ในพระราชูปถัมภ์ สำนักงานคณะกรรมการอุดมศึกษา

ข่าวประชาสัมพันธ์

1. สำหรับผู้ที่ต้องการแบบตอบรับสามารถติดต่อได้ที่ คุณยุวธิดา หองสุทธา สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม 85 ม.3 ต.นครปฐม อ.เมือง จ.นครปฐม 73000 หมายเลข โทรศัพท์ 034-261053

2. ขณะนี้ผลการพิจารณาบทความได้ส่งไปให้ท่านผู้เขียนเรียบร้อยแล้ว ท่านสามารถตรวจสอบผลได้ทาง E-Mail ที่ลงทะเบียนไว้

3. สำหรับการนำเสนอแบบบรรยาย ให้ผู้นำเสนอเตรียมไฟล์ในรูปแบบ MS PowerPoint ใช้เวลาในการบรรยาย 10 นาที และตอบข้อซักถาม 5 นาที

4. ดาวน์โหลด ศูจิบัตร <<DOWNLOAD>>

หน้าหลัก

published by admin on ศ, 04/15/2011 - 11:51

การประชุมวิชาการระดับชาติ
มหาวิทยาลัยราชภัฏนครปฐม ครั้งที่3
The 3rd NPRU National Conference 2011
Rajabhatnakhonpathom I Nakhonpathom I Thailand I 10-11 Aug 2011

หลักการและเหตุผล :

งานประชุมวิชาการ The 3rd NPRU NATIONAL CONFERENCE เป็นเวทีสำหรับการแลกเปลี่ยนทางวิชาการระหว่างอาจารย์ นักวิจัย และนักศึกษาระดับบัณฑิตศึกษาที่นำเสนอแลกเปลี่ยนความรู้ทางวิชาการจากงานวิจัย เพื่อให้เกิดเครือข่ายความร่วมมือทางวิชาการระดับชาติในสหสาขาวิชาทั้ง วิทยาศาสตร์ วิศวกรรมศาสตร์ พยาบาลศาสตร์ ครุศาสตร์ วิทยาการจัดการ มนุษยศาสตร์ และสังคมศาสตร์ บทความวิจัยที่ส่งมาจะใช้วิธีการตรวจสอบโดยผู้เชี่ยวชาญในสาขานั้นจำนวน 3 ท่าน เพื่อคัดกรอง และตรวจแก้ไขให้บทความวิจัยที่ได้รับคัดเลือกในการนำเสนอทั้งในรูปแบบปากเปล่า (Oral Presentation) และโปสเตอร์มีคุณภาพ

นอกจากนั้นบทความวิจัยที่ได้รับคัดเลือกให้นำเสนอในงานประชุมวิชาการ จะได้รับการตีพิมพ์ ในเอกสารประกอบการประชุมวิชาการ (Proceeding Book of the 3rd NPRU NATIONAL CONFERENCE) พร้อมค้วยการบันทึกบทความวิจัยฉบับเต็ม ในรูปแบบซีดีรอมไฟล์ในแผ่นดิสก์ และส่งไปยังห้องสมุดของมหาวิทยาลัยทั่วประเทศ หรือหน่วยงานทางวิชาการที่เกี่ยวข้อง เพื่อใช้ในการอ้างอิงหรือเผยแพร่ความรู้ของงานวิจัยเหล่านี้ต่อไป จากเหตุผลที่กล่าวมาข้างต้น สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม จึงขอเสนอ โครงการประชุมวิชาการ มหาวิทยาลัยราชภัฏนครปฐม ครั้งที่ 3 ภายใต้แนวความคิดของการจัดประชุมวิชาการ **"75 ปี จากโรงเรียนสตรีศึกษัดครูสู่มหาวิทยาลัยเพื่อการพัฒนาท้องถิ่นที่ยั่งยืน"** ระหว่างวันที่ 10 – 11 สิงหาคม 2554 ณ มหาวิทยาลัยราชภัฏนครปฐม

กำหนดการ :

- **เปิดรับบทความ** วันนี้ - 30 มิถุนายน 2554
- **แจ้งผลการพิจารณาบทความ** สิ้นเดือน กรกฎาคม 2554
- **ส่งบทความแก้ไขฉบับสมบูรณ์** 31 กรกฎาคม 2554
- **นำเสนอบทความ** 10 - 11 สิงหาคม 2554

Download Paper Template :

ดาวน์โหลด (Doc), ดาวน์โหลด (PDF), Poster Template

*** แก้ไขการเขียนบทคัดย่อ ความยาวไม่ควรเกินครึ่งหน้ากระดาษ

การส่งบทความสามารถส่งได้โดยตรง โดยระบุหัวข้อบทความวิจัยที่จะส่งให้ชัดเจนและส่งมาที่ research.nprw@gmail.com

กลุ่มหัวข้อที่เปิดรับบทความวิจัย :

- 1. การวิจัยเพื่อท้องถิ่น - รศ.ชลิรัตน์ พยอมแย้ม
- 2. วิทยาศาสตร์ และเทคโนโลยี
- ฟิสิกส์ - อ.ดร.จักรพงษ์ แก้วขาว
- คอมพิวเตอร์ - อ.ดร.สหจน์ เฮงพระพรหม
- สาธารณสุข - อ.ดร.สาโรจน์ เทพรมณี
- วิศวกรรมไมโครคอมพาคอมและวิศวกรรมอิเล็กทรอนิกส์ - รศ.ดร.ปิยะ โควินท์กวีวัฒน์
- ชีววิทยา เกษตร และอาหาร - อ.ดร.พิทักษ์พงศ์ ป้อมปราณี
- เคมี - อ.ดร.สมปอง ทองงามดี
- 3. ครุศาสตร์
- หลักสูตรและการสอน - อ.จำรัส อินทคาคาพร, อ.เอกชัย โวยโสภี
- วิจัยการศึกษา - อ.ดร.จันทิมา แสงเลิศอรัย
- บริหารการศึกษา - อ.ดร.ดวงใจ ช่วยตระกูล
- 4. มนุษยศาสตร์และสังคมศาสตร์
- รัฐประศาสนศาสตร์และนิติศาสตร์ - ผศ.ดร.กัณฑ์ธีมา นิลทองคำ
- ศิลปะ - อ.ทักกาล ต้นทองคำ
- มนุษยศาสตร์ และสังคมศาสตร์ - อ.สมศักดิ์ ลมสิริพงศ์
- การท่องเที่ยว - อ.ดร.นิพน เชื่อเมืองพาน
- 5. วิทยาการจัดการ
- บริหารธุรกิจและการจัดการโลจิสติกส์ - อ.ดร.แสงแข บุญศิริ
- นิเทศศาสตร์ - ผศ.ดร.กฤติยา วริโยธิน
- การบัญชี การเงิน และการธนาคาร - อ.พิมพ์นารา พิบูลวิรกานต์
- 6. พยาบาลศาสตร์ - รศ.ดร.สนตรา ตะบุนหงศ์

ดาวน์โหลด ศูจิบัตร

ดาวน์โหลด




แนวคิดในการค้นคืนข้อมูลเชิงความหมายด้านเทคโนโลยีสารสนเทศภาษาไทย

A Conceptual Framework for Thai IT Data Retrieval based on Semantic

โกญจนพงษ์ ทองเพชร[1] และ คะชา ชาญศิลป์[2]

[1]สาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี อำเภอเมือง นครราชสีมา
konjanapong@gmail.com

[2]สาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี อำเภอเมือง นครราชสีมา
kacha@sut.ac.th

บทคัดย่อ

ข่าวสารข้อมูลสารสนเทศต่างๆ ในรูปแบบอิเล็กทรอนิกส์มีเพิ่มขึ้นมากมายในปัจจุบัน จึงได้มีการพัฒนาระบบค้นคืนข้อมูลขึ้นมาใช้เพื่อให้ได้ข้อมูลที่ถูกต้องที่สุด แต่ในการค้นคืนข้อมูลในภาษาไทยนั้นยังไม่มีการพัฒนาให้สามารถค้นคืนผลลัพธ์ได้ดีเท่าที่ควร เนื่องจากรูปแบบการเขียนภาษาไทยนั้นไม่มีจุดสิ้นสุดคำที่แน่นอน ทำให้การค้นหาแบบการอ้างอิงคำหลัก (Keyword-Base) **ทำได้ไม่ดีนัก จากปัญหานี้ผู้วิจัยจึงได้เสนอ** แนวคิดในการนำเทคโนโลยีเชิงความหมาย (Semantic Technology) **มาใช้งานในการพัฒนาระบบค้นคืน** ข้อมูลภาษาไทยและใช้ออนโทโลจี (Ontology) **เข้ามาอธิบายความสัมพันธ์ของคำต่างๆในภาษาไทย ซึ่งมีการ** ทำงานในลักษณะของเว็บแอพพลิเคชัน

คำสำคัญ: ออนโทโลจี, เทคโนโลยีเชิงความหมาย, การค้นคืนข้อมูล

Abstract

Nowadays, as the amount of electronic information technology are getting more and more. Therefore, the information retrieval system was used to obtain the most accurate information. However, the data retrieval in Thai is not developed to be able to retrieve the results as good as it should. The Keyword-base is not effective enough. Because, in Thai that have no explicit boundary delimiter. Therefore, researcher has designed Thai Data Retrieval by using Semantic Technology, and use Ontology Technology to describe relationship between Thai words that work on web application.

Keyword: Ontology, Semantic Technology, Data Retrieval

1. บทนำ

ในปัจจุบันเทคโนโลยีสารสนเทศได้มีบทบาทกับการดำเนินชีวิตเป็นอย่างมาก ซึ่งส่งผลให้มีข้อมูลข่าวสารต่างๆ เพิ่มจำนวนมากขึ้นเรื่อยๆ โดยเฉพาะข้อมูลด้านเทคโนโลยีสารสนเทศ ทำให้เกิดปัญหาในการพิจารณาข้อมูลซึ่งอาจไม่ตรงกับความต้องการของผู้ใช้งาน จึงเกิดการพัฒนาระบบช่วยบริการค้นคืนข้อมูล ซึ่งทำให้การค้นคืนข้อมูลเป็นไปด้วยความรวดเร็วและถูกต้องขึ้น แต่ในการค้นคืนข้อมูลในภาษาไทยนั้น ผลลัพธ์ที่ได้จากการค้นคืนข้อมูลมักไม่ตรงตามที่ผู้ใช้งานต้องการ ปัญหาที่เกิดขึ้นนั้นเนื่องมาจากเทคนิคในการค้นคืนข้อมูลส่วนใหญ่ ตั้งอยู่บนพื้นฐานของการค้นคืนข้อมูลที่สอดคล้องหรือเหมือนกับคำหลัก (Keyword Matching) ทำให้

การแสดงผลลัพธ์ในการค้นคืนข้อมูลผิดพลาด เนื่องจากภาษาไทยมีรูปแบบการเขียนซับซ้อน อีกทั้งคำบางคำมีได้หลายความหมาย ผู้วิจัยจึงเกิดแนวคิดในการพัฒนาระบบค้นคืนข้อมูลที่แก้ปัญหาดังกล่าว ซึ่งทำให้การค้นคืนข้อมูลเป็นไปได้สะดวกรวดเร็วยิ่งขึ้น โดยนำเทคโนโลยีเชิงความหมายเข้ามาช่วยในการแปลความจากคำหลักและใช้ออนโทโลจีที่สร้างขึ้นมาจากการจัดความสัมพันธ์ของกลุ่มคำในด้านเทคโนโลยีสารสนเทศ

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 เทคโนโลยีเชิงความหมาย (Semantic Technology)

เทคโนโลยีเชิงความหมาย [6] ถูกพัฒนาขึ้นเพื่อทำให้คอมพิวเตอร์และแอพพลิเคชัน สามารถเข้าใจข้อมูลที่สอดคล้องกับความเข้าใจของมนุษย์เพื่อสามารถแลกเปลี่ยนข้อมูลที่สนใจและนำไปประมวลผลต่อไปได้ถูกต้อง การทำให้คอมพิวเตอร์สามารถเข้าใจความหมายของคำและแนวความคิดแบบเดียวกับมนุษย์นั้น จะต้องจัดการข้อมูลที่สนใจในลักษณะของการเชื่อมโยงความสัมพันธ์ของข้อมูลในระดับเมตะดาตา (Metadata) ทำให้เครื่องคอมพิวเตอร์สามารถเข้าใจความหมายของข้อมูลต่างๆ ได้เช่น เป็นข้อมูลที่เกี่ยวข้องกับอะไร มาจากส่วนไหนของชุดข้อมูล เป็นต้น ซึ่งจากการทำงานในระบบต่างๆ ของคอมพิวเตอร์ที่สามารถเข้าใจความหมายได้ตรงกับมนุษย์นั้น จะทำให้ได้ผลลัพธ์ที่สามารถลดปริมาณและระยะเวลาในการทำงานให้น้อยลงได้

2.2 ออนโทโลจี (Ontology)

ออนโทโลจี [1] หมายถึงวิธีการบรรยายแนวความคิดตามขอบเขตที่ต้องการหรือข้อกำหนดที่เกี่ยวกับแนวคิด ซึ่งเป็นการสร้างโครงสร้างฐานความรู้ทางด้านใดด้านหนึ่งหรือขอบเขตใดขอบเขตหนึ่งที่มีแนวคิดและความเข้าใจที่ตรงกัน อีกทั้งยังใช้ในการอธิบายความหมายของสิ่งต่าง ๆ และจัดหมวดหมู่ของข้อมูลได้ในขอบเขตความสนใจหนึ่ง ๆ โดยแสดงออกมาในรูปของคำสามัญเพื่อให้เข้าใจและนำไปใช้ได้ร่วมกันระหว่างคอมพิวเตอร์และมนุษย์ได้ ในภาษาธรรมชาติสามารถแสดงได้โดยใช้คำและประโยค เพื่อแสดงความสัมพันธ์ระหว่างคำศัพท์เหล่านั้น ส่วนการนำไปใช้ในด้านคอมพิวเตอร์จะแสดงในรูปแบบของระบบสัญลักษณ์ (Notation) เช่น คลาส (Class) อินสแตนซ์ (Instance) ความสัมพันธ์ (Relation) คุณสมบัติ (Property) กฎ (Rule) เป็นต้น โดยใช้ภาษาสำหรับแสดงความรู้ (Knowledge Representation Language) ซึ่งมีความชัดเจนและเที่ยงตรงมากกว่าคำศัพท์และประโยคในภาษาธรรมชาติ

2.3 งานวิจัยที่เกี่ยวข้อง

ธนกร หวังพิพัฒน์วงศ์ [8] ได้นำเสนอรูปแบบการค้นหารูปภาพสถานที่ท่องเที่ยวในประเทศไทย โดยนำหลักการเชิงความหมายมาช่วยในการพัฒนาระบบสืบค้น ซึ่งได้รวบรวมคำบรรยายภาพของแต่ละภาพมาจัดลำดับความสัมพันธ์กัน และได้นำฐานข้อมูลออนโทโลจี WordNet [3] เข้ามาช่วยในการอ้างอิงความสัมพันธ์ของคำศัพท์ภาษาอังกฤษ ทำให้การสืบค้นแม่นยำมากขึ้น

Yi Jin [7] **ได้ทำการศึกษาและนำเสนอการนำอัลกอริทึม TFIDF เข้ามาช่วยในการประมวลผลในการ**แยกคำหรือการตัดคำ ซึ่งเป็นการนำหลักการทางสถิติเข้ามาช่วย โดยการทำงานจะใช้การจัดทำดัชนีชี้วัดค่าความสำคัญของข้อมูล ซึ่งช่วยประเมินความสำคัญของข้อมูลในการนำมาแสดงผล เพื่อเพิ่มความถูกต้องในการสืบค้น ซึ่งแสดงให้เห็นว่าผลลัพธ์ของการค้นคืนข้อมูลแสดงได้ถูกต้องมากขึ้น แต่เนื่องจากงานวิจัยนี้ได้ใช้อัลกอริทึมที่อ้างอิงค่าจากจำนวนคำหลักของเอกสาร ผู้วิจัยจึงเสนอแนวทางในการพัฒนาเพื่อให้ได้ผลลัพธ์ที่ถูกต้อง

มากขึ้น โดยการวิเคราะห์ค่าของคำหลักที่มีผลต่อการสืบค้นข้อมูลตามความเป็นจริงโดยไม่อ้างอิงจากจำนวนคำที่มีอยู่ในเอกสาร เพื่อให้ได้ผลลัพธ์ของการสืบค้นที่ถูกต้องมากขึ้น

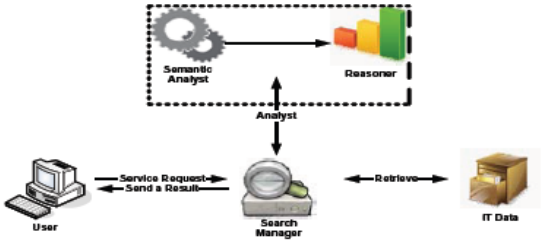
Junaidah Mohamed [2] ได้นำเสนอแนวคิดเกี่ยวกับหลักการของการพัฒนาระบบบริการสืบค้นโดยใช้หลักการเว็บเชิงความหมายมาช่วยพัฒนา ซึ่งได้อธิบายภาพรวมของหลักการและนำเสนอเทคโนโลยีที่จำเป็นต่อการพัฒนาระบบ อีกทั้งยังได้นำเสนอเว็บไซต์ที่ให้บริการสืบค้นสารสนเทศซึ่งใช้หลักการเชิงความหมาย สำหรับผู้ที่สนใจสามารถเข้าไปทดลองใช้บริการสืบค้นได้ พร้อมทั้งอธิบายและแสดงการเปรียบเทียบข้อดีข้อเสียของการค้นคืนแบบต่างๆ

Reza Hemayati [5] ได้นำหลักการจัดกลุ่มของคำมาช่วยเพิ่มประสิทธิภาพของผลลัพธ์ในการสืบค้นข้อมูลสารสนเทศ โดยนำเสนอ SRR Grouping Algorithm สองรูปแบบเปรียบเทียบกันคือ Largest Frequency of Use (LF) และ Largest Category (LC) ซึ่งระบบได้ทำการพัฒนาด้วยภาษาจาวาและฐานข้อมูลออนโทโลจี WordNet ในการอ้างอิงความสัมพันธ์ของคำ จากผลการทดสอบแสดงให้เห็นว่าการทำอัลกอริทึมทั้งสองรูปแบบมาใช้ในการพัฒนาระบบสืบค้นข้อมูลสารสนเทศ ทำให้ได้ผลลัพธ์ของการสืบค้นข้อมูลได้แม่นยำเฉลี่ยสูงสุด 93 เปอร์เซนต์

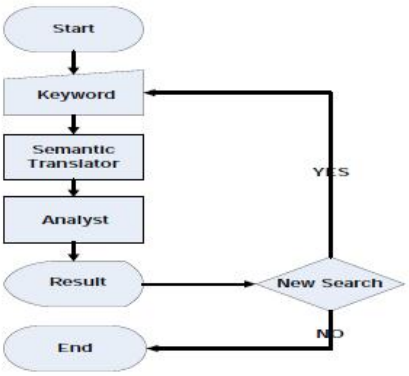
3. การออกแบบและพัฒนาระบบ

3.1 การออกแบบระบบ

ในกระบวนการทำงานของระบบ จะมีการตรวจสอบข้อมูลโดยระบบจะทำการสร้างดัชนี (Index) ของเอกสารแต่ละชนิดที่ผู้ใช้งานต้องการขึ้นมา แล้วทำการเก็บข้อมูลไว้ในระบบ จากนั้นระบบจะนำเอาข้อมูลมาสร้างเป็นฐานความรู้เก็บไว้ โดยใช้ความสัมพันธ์กันระหว่างข้อมูลต่างๆ ที่ได้ป้อนเข้าสู่ระบบ เมื่อผู้ใช้งานต้องการค้นหาข้อมูล ระบบค้นคืนข้อมูลจะทำการค้นคืนข้อมูลโดยกระบวนการเชิงความหมายและแสดงผลลัพธ์ให้ผู้ใช้งาน ดังแสดงในภาพที่ 1 และภาพที่ 2



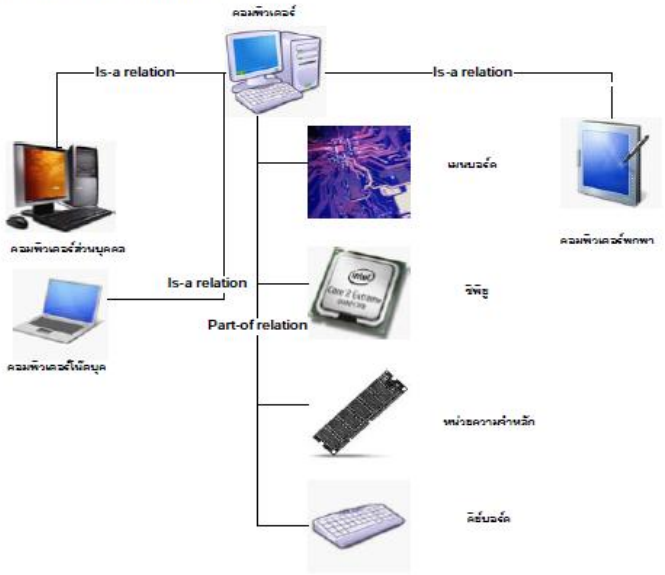
ภาพที่ 1: แสดงภาพรวมการทำงานของระบบ



ภาพที่ 2: แสดงขั้นตอนการทำงานของระบบ

3.2 การพัฒนาระบบงาน

ในการพัฒนาระบบงานนั้น ผู้วิจัยได้เลือกใช้การพัฒนาให้อยู่ในรูปแบบของเว็บแอพพลิเคชัน โดยใช้เทคโนโลยีภาษาจาวาและส่วนเสริมลูซีน 3.0 (Lucene3.0) ในการพัฒนา นอกจากนั้นในการสร้างออนโทโลจีเพื่อใช้ในการอธิบายความหมายของข้อมูลได้มีการนำโปรแกรม Protégé [4] มาใช้ในการออกแบบโครงสร้างความสัมพันธ์ของข้อมูลคำศัพท์ต่างๆ และใช้ภาษาโอดับบลิวแอล (OWL) ในการสร้างออนโทโลจีของข้อมูล **ซึ่งมีตัวอย่าง** โครงสร้างของออนโทโลจีดังแสดงในภาพที่ 3



ภาพที่ 3: แสดงรูปแบบความสัมพันธ์ของออนโทโลจีภาษาไทย

จากภาพที่ 3 **ได้แสดงคลาสของคำว่าคอมพิวเตอร์และแสดงสัมพันธ์ย่อยต่างๆ โดยมีตัวอย่างรายละเอียดดัง** แสดงในตารางที่ 1

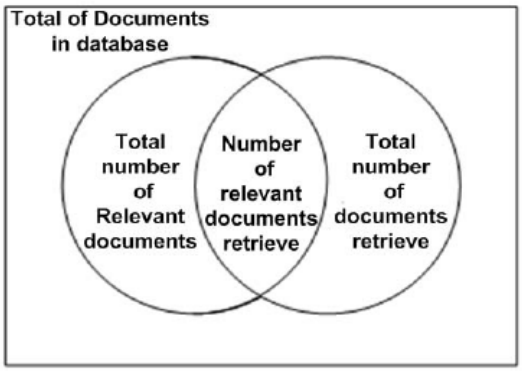
ID	คำศัพท์	คำอธิบายคำศัพท์	คำศัพท์ที่เกี่ยวข้อง	ความสัมพันธ์แบบ Superclass	ความสัมพันธ์แบบ Subclass	ความสัมพันธ์แบบ Sibling
1	คอมพิวเตอร์		เมนบอร์ด, **ซีพียู, หน่วยความจำหลัก,** คีย์บอร์ด		คอมพิวเตอร์ส่วนบุคคล, โน๊ตบุค, คอมพิวเตอร์พกพา	
2	โน๊ตบุค		เมนบอร์ด, **ซีพียู, หน่วยความจำหลัก,** คีย์บอร์ด	คอมพิวเตอร์		คอมพิวเตอร์ส่วนบุคคล, โน๊ตบุค, คอมพิวเตอร์พกพา

ตารางที่ 1: แสดงตัวอย่างรายละเอียดฐานข้อมูลออนโทโลจี

จากตารางที่ 1 **จะเห็นได้ว่าในแต่ละคำของออนโทโลจีจะมีการอธิบายความหมายและความสัมพันธ์กับคำ** อื่นๆ ไว้อย่างละเอียด ทำให้เมื่อนำไปใช้งานในการค้นคืนข้อมูลจะแสดงผลลัพธ์ในการค้นคืนข้อมูลได้ละเอียดและถูกต้องตามความหมายมากขึ้น โดยคำส่วนใหญ่ที่ผู้วิจัยจะนำมาทำการสร้างฐานข้อมูลออนโทโลจีแบ่งเป็นกลุ่มหลักได้สองกลุ่ม กลุ่มแรกคือกลุ่มคำที่เขียนตามเสียงอ่านของภาษาอังกฤษหรือที่มักเรียกกันว่าคำทับศัพท์ เนื่องด้วยบางครั้งการเขียนอาจจะเขียนไม่เหมือนกันขึ้นอยู่กับความเข้าใจของผู้เขียนแต่ละท่าน ทำให้อาจเกิดความสับสนในการค้นคืนโดยใช้คำหลักภาษาไทย และอีกกลุ่มหนึ่งคือกลุ่มคำที่ใช้กันทั่วไปซึ่งจะรวบรวมกลุ่มคำเหล่านี้โดยการสุ่มเอกสารงานวิจัยที่เกี่ยวกับด้านเทคโนโลยีสารสนเทศมาจำนวนหนึ่งแล้วคัดเลือกคำศัพท์ด้านเทคโนโลยีสารสนเทศที่มีการใช้งานมาจัดเป็นกลุ่มคำ โดยกลุ่มคำเหล่านี้จะเป็นกลุ่มคำที่มักมีการใช้งานบ่อยและผู้ใช้งานสามารถเข้าใจรูปแบบการเขียนได้ตรงกันแต่บางคำอาจมีได้หลายความหมายจึงได้มีการนำมาจัดความสัมพันธ์ในรูปแบบออนโทโลจีเพื่อแสดงความ สัมพันธ์ของคำศัพท์ ซึ่งเมื่อนำไปใช้งานในการค้นคืนข้อมูลจะทำให้ผลลัพธ์ที่ได้จากการค้นคืนมีความถูกต้องมากขึ้น

4. การทดสอบและประเมินผล

จากแนวคิดในการค้นคืนข้อมูลเชิงความหมายด้านเทคโนโลยีสารสนเทศภาษาไทยที่ผู้วิจัยได้นำเสนอในตอนต้นนั้น สามารถทำการทดสอบและประเมินผลได้โดยใช้การวัดค่าความแม่นยำ (Precision) **ค่าความถูก**ต้อง (Recall) โดยจะแสดงดังภาพที่ 4



ภาพที่ 4: แสดงเซตของเอกสารทั้งหมด

สมการที่ใช้หาค่าประสิทธิภาพ Precision และ Recall แสดงได้ดังนี้คือ

Precision = Number of relevant documents retrieve / Total number of documents retrieve (4-1)

จากสมการที่ 4-1 **คำนวณค่าความแม่นยำได้จากเอกสารที่ค้นคืนทั้งหมดหารด้วยจำนวนเอกสารทั้งหมดที่มี**ความเกี่ยวข้อง

Recall = Number of relevant documents retrieve / Total number of relevant documents (4-2)

จากสมการที่ 4-2 **คำนวณค่าความถูกต้องได้จากเอกสารที่ค้นคืนและเกี่ยวข้องทั้งหมดหารด้วยจำนวน**เอกสารทั้งหมดที่ค้นคืนได้

ในการทดสอบและประเมินผลความแม่นยำผู้วิจัยจะใช้เอกสารที่เกี่ยวข้องจำนวน 200 **เอกสาร ซึ่ง**เอกสารทั้งหมดเป็นเอกสารภาษาไทยที่อยู่ในรูปแบบข้อมูล PDF File Format และผู้วิจัยจะทำการสุ่มทดลองค้นคืนข้อมูลโดยใช้คำหลักต่างกัน 20 **คำ ซึ่งเป็นคำภาษาไทยที่เกี่ยวข้องกับด้านเทคโนโลยีสารสนเทศ ผลลัทธ์ที่**ได้จะถูกคำนวณค่าความถูกต้องตามสมการที่ 4-1 และ 4-2 ในส่วนของเกณฑ์ในการประเมินประสิทธิภาพของระบบคือค่าความแม่นยำ (Precision) **และค่าความถูกต้อง (Recall) ต้องมีค่ามากกว่า 60 % จึงจะถือว่าระบบ**ที่พัฒนามีประสิทธิภาพดี

5. สรุปผลและข้อเสนอแนะ

5.1 สรุปผลการดำเนินงาน

งานวิจัยนี้เสนอแนวคิดในการแก้ปัญหาการค้นคืนข้อมูลเชิงความหมายด้านเทคโนโลยีสารสนเทศภาษาไทยโดยใช้ออนโทโลจีเพื่อลดเวลาในการคัดเลือกข้อมูลที่ต้องการจากผลลัพธ์ที่ได้มาของผู้ใช้งานในการค้น

คืนข้อมูลและเพิ่มความถูกต้องแม่นยำในการค้นคืนข้อมูลซึ่งใช้เทคโนโลยีเชิงความหมายในการวิเคราะห์คำหลักของการค้นคืน ร่วมกับฐานข้อมูลออนโทโลจีที่พัฒนาขึ้น โดยผลลัพธ์ที่ได้จะถูกวัดประสิทธิภาพด้วยค่า Precision และค่า Recall ซึ่งจะมีการรายงานผลในโอกาสต่อไป

5.2 ข้อเสนอแนะ

5.2.1 สร้างออนโทโลจีจากแหล่งข้อมูลที่สนใจมากกว่าแหล่งข้อมูลเดียว แล้วนำออนโทโลจีที่ได้มาทำการรวมเข้าด้วยกัน จะทำให้เกิดความหลากหลายในการค้นคืนข้อมูลที่มากขึ้น

5.2.2 นำไปใช้ในการพัฒนาระบบค้นคืนข้อมูลด้านอื่นๆ **เช่นข้อมูลวิทยานิพนธ์ของมหาวิทยาลัยต่างๆ** เป็นต้น

6. เอกสารอ้างอิง

[1] Alexander M. and Steffen S., "Ontology Learning for the Semantic Web", University of Karlsruhe, Karlsruhe Germany, 2001.

[2] Junaidah Mohamed kassim and Mahathir Rahmany, "Introduction to Semantic Search Engine", Electrical Engineering and Informatics ICEEI '09, 5 August 2009.

[3] J. Morato, M.A. Marzal, J. Llorens and J. Moreiro, "WordNet Application", GWC 2004, pp. 270-278.

[4] Knublauch H., Ray W. Fergerson, Natalya F. Noy and Mark A., "The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications", Proceedings of the Semantic Web-ISWC 2004, 2004.

[5] Reza Hemayati, Weiyi Meng and Clement Yu., "Semantic-Based Grouping of Search Engine Results Using WordNet", APWeb/WAIM 2007.

[6] Tim Berners-Lee, James Hendler and Ora Lassila., "The semantic Web", Scientific American Magazine May 2001.

[7] Yi Jin, Zhuying Lin and Hongwei Lin., "The Research of Search Engine Based on Semantic Web", International Symposium on Intelligent Information Technology Application Workshops 2008.

[8] ธนกร หวังพิพัฒน์วงศ์, อานนท์ ไกรเสวกวิสัย และสราวุธิ ราษฎร์นิยม, "ระบบค้นหารูปภาพโดยใช้หลักการเว็บเชิงความหมาย", มหาวิทยาลัยกรุงเทพ, กรุงเทพฯ, 2009.

ประวัติผู้เขียน

นายโกญจนพงษ์ ทองเพชร เกิดเมื่อวันที่ 26 ตุลาคม 2527 ที่อำเภอเมือง จังหวัดตราด จบการศึกษาระดับประถมศึกษาที่โรงเรียนบึงพญาปราบ ระดับมัธยมศึกษาตอนต้นและมัธยมศึกษาตอนปลายที่โรงเรียนราชสีมาวิทยาลัย จากนั้นได้สอบเข้าเรียนระดับอุดมศึกษาที่มหาวิทยาลัยเทคโนโลยีสุรนารี จังหวัดนครราชสีมา ในสำนักวิชาวิศวกรรมศาสตร์ เมื่อปี2546 จบการศึกษาระดับปริญญาตรีในสาขาวิชาวิศวกรรมคอมพิวเตอร์ เมื่อปี 2550 หลังจบการศึกษาได้ทำงานเป็นพนักงานฝ่ายพัฒนาซอฟต์แวร์ ในสัญญาจ้าง 6 เดือน ที่สำนักงานส่งเสริมอุตสาหกรรมซอฟต์แวร์แห่งชาติ องค์การมหาชน สาขาเชียงใหม่ หลังจากนั้นได้เข้าเป็นพนักงานประจำในตำแหน่งพนักงานฝ่ายพัฒนาเว็บไซต์ที่บริษัทฟินไลน์ จำกัด เมื่อปี 2551 ทำหน้าที่ดูแลและพัฒนาระบบเว็บไซต์ฝ่ายเช่าซื้อของ ธนาคารธนชาต จำกัด มหาชน

ในปี 2548 ได้เข้ารับการอบรมสอบวัดความรู้ด้านภาษาจาวาที่มหาวิทยาลัยขอนแก่น จัดโดยสำนักงานส่งเสริมอุตสาหกรรมซอฟต์แวร์แห่งชาติ องค์การมหาชน สาขาขอนแก่น และสอบผ่านได้ใบ Certified Java 1.4 เมื่อเดือนมีนาคม 2552 ได้ลาออกจากบริษัทฟินไลน์ จำกัด เพื่อเข้าศึกษาต่อในระดับปริญญาโท สาขาวิศวกรรมคอมพิวเตอร์ สำนักวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

ผลงานวิจัย : ได้นำเสนอบทความเข้าร่วมในการประชุม NPRU National Conference 2011 เมื่อวันที่ 10 – 11 สิงหาคม 2554 ที่มหาวิทยาลัยราชภัฏนครปฐม โดยได้เสนอบทความเรื่อง แนวคิดในการค้นคืนข้อมูลเชิงความหมายด้านเทคโนโลยีสารสนเทศภาษาไทย