โครงร่างโครงงานวิศวกรรมคอมพิวเตอร์

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์

มหาวิทยาลัยเกษตรศาสตร์

เรื่อง

การวิเคราะห์คุณลักษณะของเว็บไทย
Analysis of the Thai web characteristics

โดย

นายวิวัฒน์ นิ่มสนิทนันท์ รหัส 5310504001

พ.ศ. 2556

การวิเคราะห์คุณลักษณะของเว็บไทย
Analysis of the Thai web characteristics

โดย
นาย วิวัฒน์ นิ่มสนิทนันท์
รหัสประจำตัว 5310504001

โครงงานวิศวกรรมคอมพิวเตอร์
ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์
มหาวิทยาลัยเกษตรศาสตร์

ตามหลักสูตร
วิศวกรรมศาสตรบัณฑิต
สาขาวิศวกรรมคอมพิวเตอร์

ได้รับการพิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษา ..... วันที่.....เดือน .....พ.ศ.....
(ผศ.ดร. อานนท์ รุ่งสว่าง)
กรรมการ ..... วันที่.....เดือน .....พ.ศ.....
(ผศ.ดร. ภุชงค์ อุทโยภาส)

วิวัฒน์ นิ่มสนิทนันท์ ปีการศึกษา 2556
การวิเคราะห์คุณลักษณะของเว็บไทย
ปริญญาวิศวกรรมศาสตรบัณฑิต (สาขาวิศวกรรมคอมพิวเตอร์) ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

บทคัดย่อ

การพัฒนาระบบเสิร์จเอ็นจินเฉพาะภายในประเทศ มีวัตถุประสงค์เพื่อตอบสนองความต้องการการค้นคืนที่ตรงตามความต้องการผู้ใช้ในประเทศนั้น เสิร์จเอ็นจินที่ประสบความสำเร็จสูงสุดของประเทศจีนคือ Baidu ซึ่งพบว่าคนจีนส่วนใหญ่นิยมใช้เสิร์จเอ็นจินของ Baidu มากกว่าเสิร์จเอ็นจินของ Google ซึ่งสิ่งจำเป็นในการพัฒนาเสิร์จเอ็นจินเฉพาะภายในประเทศอย่างหนึ่งคือ การสังเกตและวิเคราะห์คุณลักษณะของเว็บประเทศที่สนใจ ดังนั้นโครงงานนี้ จะทำการสกัดหาคุณลักษณะของเว็บไทยแล้ววิเคราะห์ทั้งในด้านเนื้อหาและโครงสร้างการเชื่อมโยงของเว็บเพจเหล่านั้น จากการทดลองสร้างระบบสืบค้นบนฐานข้อมูลเว็บเพจภาษาไทยที่เก็บรวบรวมจากอินเตอร์เน็ตเป็นจำนวนประมาณ 4ล้านเว็บเพจ พบว่าการจัดหมวดหมู่เว็บเพจตามคุณลักษณะที่วิเคราะห์ได้สามารถนำมาช่วยในการลดความกำกวมของความหมายของคำค้นได้เป็นอย่างดี

คำสำคัญ คุณลักษณะเว็บไทย การจัดกลุ่มเว็บเพจ การพัฒนาเสิร์จเอ็นจิน

เลขที่เอกสารอ้างอิงภาควิชา .....

Wiwat Nimsanitnan Academic Year 2013
Analysis of the Thai web characteristics
Bachelor degree in Engineering, Department of Computer Engineering
Faculty of Engineering, Kasetsart University

Abstract

Web search system development within the country aims to meet the retrieval to serve the users' needs in that country. The highest successful web search engine of China is Baidu, in which we found that most Chinese people quite prefer Baidu than Google. The essential efford for developing search engine used within the country is an observation and analysis of characteristics of web pages resided in that country. In this engineering project, we first extract features of Thai web pages and then analyze their both content and links. The results of experiments conducted on retrieval system over 4 million Thai web pages collected from the Internet show that classification by analyzed features can help to reduce the ambiguity of the meaning of query terms in retrival process.

Keywords: characteristic of Thai web page, web page classification, search engine development

Department Reference No .....

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบคุณอาจารย์ที่ปรึกษา ผศ.ดร.อานนท์ รุ่งสว่าง และ ดร.บัณฑิต มนัสเกษมศักดิ์ ที่เป็นผู้ให้คำปรึกษา ให้แนวคิดในการทำงานวิจัย รวมถึงการปลูกฝังจิตสำนึกต่อส่วนรวม และการแบ่งปัน อีกทั้งยังคอยดูแลข้าพเจ้าในช่วงการเรียนในมหาวิทยาลัยเกษตรศาสตร์มาเป็นตลอดระยะเวลา 4 ปีที่ผ่านมา ขอขอบคุณเพื่อน ๆ นิสิตภาควิชาวิศวกรรมคอมพิวเตอร์รุ่นที่ 24 ที่คอยช่วยเหลือและเป็นกำลังใจให้แก่ข้าพเจ้าเสมอมา

วิวัฒน์ นิ่มสนิทนันท์
ผู้จัดทำ

สารบัญ

สารบัญภาพ

สารบัญตาราง

1 บทนำ

คุณลักษณะของข้อมูลเว็บแต่ละประเทศ สามารถนำมาใช้วิเคราะห์เพื่อพัฒนาระบบสืบค้นข้อมูลของชาติที่มีประสิทธิภาพได้ ยกตัวอย่างเช่น ในเว็บคราวเลอร์เจาะจงภาษา จะมีพารามิเตอร์บางอย่างที่เราต้องศึกษาคุณลักษณะของเว็บกราฟภาษาไทย เว็บคราวเลอร์นั้นถึงจะใช้ทรัพยากรในการเก็บเว็บเพจได้อย่างคุ้มค่าหากเราสามารถวิเคราะห์หาคุณลักษณะบางอย่างของเว็บไทย เช่น การเชื่อมโยงกันระหว่างเว็บเพจ 2 กลุ่ม เราก็อาจจะสามารถไปใช้สร้างระบบ Link Suggestion หรือแนะนำสินค้าบริการต่างๆ ได้ด้วยเช่นกัน นอกจากนี้การศึกษาคุณลักษณะของเว็บไทยยังอาจจะใช้คุณลักษณะที่ศึกษามาได้เหล่านั้นมาทำนายคุณลักษณะของเว็บกราฟทั้งโลกโดยรวมได้

ที่ผ่านมามีนักวิจัยหลายท่าน ได้ศึกษาคุณลักษณะเว็บกราฟของประเทศตัวเอง อาทิเช่น การวิเคราะห์คุณลักษณะของเว็บสเปน[1] ซึ่งทำการวิจัยเพื่อศึกษาคุณลักษณะแต่ละด้านของเว็บสเปนอย่างละเอียด รวมทั้งอธิบายพฤติกรรมบางอย่างที่พบ การวิเคราะห์คุณลักษณะของโฮสต์กราฟของประเทศไทย[5] ได้ศึกษาเกี่ยวกับลักษณะโครงสร้างเว็บกราฟในประเทศไทย และอัตราส่วนจำนวนเว็บไทยที่มีลิงก์ไปยังแต่ละ Top level domain name เป็นต้น แต่อย่างไรก็ตาม แม้ว่างานวิจัยเว็บไทยจะมีมาก่อนหน้านั้นแล้ว แต่เนื่องจากเว็บเพจนั้นมีการเติบโต และเปลี่ยนแปลงตลอดเวลาคุณลักษณะบางอย่างอาจจะเปลี่ยนตามไปด้วย

ดังนั้นผู้วิจัยจึงมีแนวคิดที่จะพัฒนาระบบเก็บรวบรวมข้อมูลเว็บประเทศไทย และศึกษาวิเคราะห์คุณลักษณะต่างๆ ของเว็บเพจประเทศไทยใหม่อีกครั้ง เพื่อให้ได้ข้อมูลเว็บไทยในปัจจุบัน

1.1 วัตถุประสงค์ของโครงงาน

1) เพื่อพัฒนาระบบเก็บรวบรวมเว็บเพจประเทศไทย
2) ศึกษาคุณลักษณะต่างๆ ที่จะสามารถใช้ปรับปรุงและพัฒนาระบบสืบค้นข้อมูลเว็บเพจภาษา

1.2 ขอบเขตของโครงงาน

เว็บเพจที่นำมาใช้ในการพิจารณาจะเลือกเฉพาะเว็บไทยเท่านั้น ซึ่งเว็บที่จัดว่าเป็นเว็บไทยต้องมีคุณสมบัติอย่างใดอย่างหนึ่งในสามข้อดังต่อไปนี้

1) เป็นเว็บเพจที่ชื่อโดเมนลงท้ายด้วย .th
2) เป็นเว็บเพจที่มี IP Address อยู่ในประเทศไทย
3) เป็นเว็บเพจที่มี คำภาษาไทยเกิน 10% ในเว็บเพจนั้น

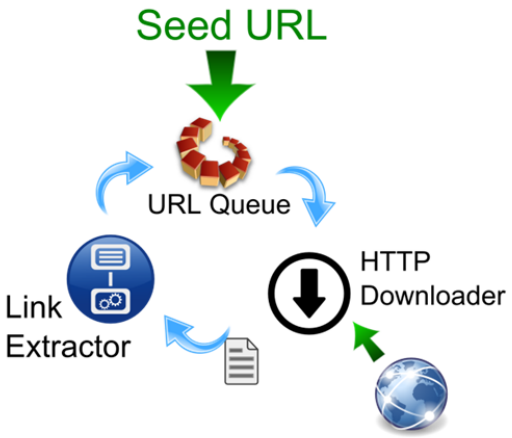
1.3 ประโยชน์ที่ได้รับ

สามารถสกัดข้อมูลคุณลักษณะเฉพาะของเว็บไทยที่จะเป็นประโยชน์ต่อการพัฒนาระบบสืบค้นข้อมูลไทยให้มีประสิทธิภาพต่อไปในอนาคต

2 ความรู้พื้นฐานและทฤษฎีที่เกี่ยวข้อง

2.1 เว็บคลาวเลอร์แบบพื้นฐาน

เว็บคลาวเลอร์คือโปรแกรมที่ใช้รวบรวมเว็บเพจอัตโนมัติ เพื่อนำเว็บเพจที่ได้นั้นมาใช้ทำระบบสืบค้นข้อมูล หรือเก็บไว้ทำเว็บอาร์ไคฟ์ (web archive) เป็นต้น



รูปที่ 1 การทำงานของเว็บคลาวเลอร์พื้นฐาน

เว็บคลาวเลอร์จะเริ่มต้นจากการรับค่ายูอาร์แอลเริ่มต้นจากผู้ใช้ แล้วเก็บยูอาร์แอลเหล่านั้นไว้ในยูอาร์แอลคิว ถัดมาก็จะนำยูอาร์แอลลำดับแรกในยูอาร์แอลคิวมาดาวน์โหลด จากนั้นนำไปสกัดลิงก์ภายในเว็บเพจ เพื่อนำลิงก์ที่ได้ไปบรรจุลงในยูอาร์แอลคิว แล้วก็เริ่มนำยูอาร์แอลลำดับถัดมาจากยูอาร์แอลคิวมาทำกระบวนการเช่นกันกับยูอาร์แอลลำดับแรก จนกระทั่งครบจำนวนที่ต้องการ หรือลิงก์ในยูอาร์แอลคิวทั้งหมดได้ถูกดาวน์โหลดไปจนหมดแล้ว ที่กล่าวมาเป็นการทำงานคร่าวๆ ของเว็บคลาวเลอร์ ซึ่งมีรายละเอียดดังต่อไปนี้

1) ยูอาร์แอลคิว

ยูอาร์แอลคิวนั้นเป็นโครงสร้างข้อมูลชนิดหนึ่ง ที่ใช้เก็บยูอาร์แอลของเว็บเพจไว้อย่างเป็นลำดับ และไม่ซ้ำกัน โดยมีหลักการคือก่อนที่จะเพิ่มยูอาร์แอลเข้าไปในยูอาร์แอลคิวนั้นต้องเช็คดูว่าไม่เคยมียูอาร์แอลดังกล่าวอยู่ในยูอาร์แอลคิวจึงจะเพิ่มยูอาร์แอลเข้าไป เพื่อรอให้เว็บคลาวเลอร์ดึงยูอาร์แอลไปดาวน์โหลดต่อ

2) ส่วนสกัดลิงก์

เว็บเพจในปัจจุบันถูกเขียนด้วยภาษาเอชทีเอ็มแอล (HTML) ซึ่งจะเป็นภาษาที่ใช้แท็ก (tag) ในการจัดโครงสร้างเว็บเพจโดยจะมีชื่อแท็กเพื่อระบุว่าโครงสร้างของเนื้อหา (Content) ส่วนนี้มีโครงสร้างแบบใด ซึ่งภายในแท็กเริ่มต้น (Start Tag) จะมีการกำหนดแอททริบิวต์ (Attribute) ภายในแท็กด้วย โดยเนื้อหาจะแทรกอยู่ระหว่างแท็กก่อนถึงแท็กสิ้นสุด (End Tag) ซึ่งการกำหนดลิงก์ให้กับเนื้อหาสามารถทำได้โดยการใส่แท็ก <a> ครอบเนื้อหาไว้ และกำหนดแอทริบิวต์ href ให้มีค่าเป็นลิงก์ที่ต้องการเชื่อมโยง

ค่าของแอททริบิวต์ href คือลิงก์ที่เราต้องการนำไป แปลงให้อยู่ในรูปยูอาร์แอลมาตรฐาน ก่อนนำไปเข้าสู่ยูอาร์แอลคิวเพื่อการดาวน์โหลดในครั้งถัดๆ ไป ของเว็บคลาวเลอร์

3) การแปลงลิงก์ให้อยู่ในรูปยูอาร์แอลมาตรฐาน (URL Normalization)

การสร้างลิงก์เชื่อมโยงไปยังหน้าเว็บเพจอื่นนั้น ผู้สร้างเว็บเพจสามารถเขียนลิงก์เชื่อมโยงไปยังเว็บเพจเป้าหมายเดียวกันแต่ใช้ยูอาร์แอลที่มีชื่อต่างกันได้ ดังนั้นผู้พัฒนาเว็บคลาวเลอร์จึงต้องแปลงยูอาร์แอลให้อยู่ในรูปมาตรฐาน (Normalize) เสียก่อน เพื่อป้องกันปัญหาในการดาวน์โหลดและเก็บเว็บเพจที่ซ้ำซ้อนตามมา ซึ่งโดยปกติการเขียนลิงก์สามารถเขียนได้ทั้งที่เป็นยูอาร์แอลสัมบูรณ์ (Absolute URL) ยูอาร์แอลสัมพัทธ์ (Relative URL) หรือในรูปพาธ (Path) ซึ่งลิงก์ที่เป็นพาธไม่สามารถส่งเข้าสู่ส่วนดาวน์โหลดได้ จึงจำเป็นต้องแปลงทุกพาธให้อยู่ในรูปยูอาร์แอลก่อน แล้วค่อยแปลงยูอาร์แอลสัมพัทธ์ให้เป็นยูอาร์แอลสัมบูรณ์ สุดท้ายก็ทำการแปลงยูอาร์แอลสัมบูรณ์ให้อยู่ในรูปมาตรฐาน มีรายละเอียดดังนี้

2.1) การแปลงลิงก์ในรูปพาธให้เป็นยูอาร์แอล แบ่งการพิจารณาเป็น 2 แบบดังนี้

2.1.1) หากพาธขึ้นต้นด้วยเครื่องหมายทับ (/) จะนำยูอาร์แอลต้นทางของเว็บเพจที่กำลังพิจารณามาตัดพาธทิ้งให้เหลือเฉพาะโปรโตคอลและโดเมน แล้วนำมาต่อด้วยพาธที่กำลังพิจารณา

2.1.2) หากพาธไม่ได้ขึ้นต้นด้วยเครื่องหมายทับ จะนำยูอาร์แอลต้นทางของเว็บเพจนั้นมาตัดชื่อไฟล์ท้ายยูอาร์แอลทิ้ง (ถ้ามี) แล้วต่อด้วยพาธที่กำลังพิจารณา

2.2) การแปลงยูอาร์แอลสัมพัทธ์ให้เป็นยูอาร์แอลสัมบูรณ์ จะตัดแฟ้มย่อย (Directory) ที่มีชื่อเป็นจุด (./) ทิ้งแล้วต่อด้วยแฟ้มลูก (Child Directory) ถัดไป และแฟ้มย่อยที่มีชื่อเป็นจุดสองจุด (../) จะทำการลบแฟ้มแม่ (Parent Directory) ที่อยู่ก่อนหน้าหนึ่งลำดับ และแฟ้มปัจจุบันทิ้งแล้วต่อด้วยแฟ้มลูก

2.3) การแปลงยูอาร์แอลสัมบูรณ์ให้อยู่ในรูปยูอาร์แอลมาตรฐาน

2.3.1) แปลงโปรโตคอลและชื่อโฮสต์ให้เป็นอักษรตัวเล็ก เช่น Http://www.Google.com จะถูกแปลงเป็น http://www.google.com

2.3.2) ลดรูปยูอาร์แอลที่เชื่อมโยงภายในหน้าเอกสารเดียวกัน เช่น http://abc.com/a.html#x จะถูกลดรูปเป็น http://abc.com/a.html

2.3.3) ถอดรหัสยูอาร์แอลในกรณีที่มีการเข้ารหัสอักขระพิเศษ เช่น http://abc.com%2Fa.html จะถูกถอดรหัสเป็น http://abc.com/a.html

2.3.4) สำหรับยูอาร์แอลที่ไม่มีพาธต่อท้าย จะเพิ่มเครื่องหมายทับ (/) ต่อท้ายยูอาร์แอลนั้น เช่น http://www.ku.ac.th จะถูกแปลงเป็น http://www.ku.ac.th/

2.2 เว็บคลาวเลอร์แบบแยกเก็บตามไซต์

เนื่องจากเว็บคลาวเลอร์พื้นฐานจัดเก็บยูอาร์แอลคิวทั้งหมดไว้ด้วยกัน และต้องเช็คการปรากฏของยูอาร์แอล ทำให้ยูอาร์แอลคิวมีขนาดใหญ่มากและเสียเวลาในการเช็คการปรากฏของยูอาร์แอลก่อนเพิ่มเข้าไปในยูอาร์แอลคิว จึงได้มีการนำเสนอเว็บคลาวเลอร์แบบแยกเก็บตามไซต์[2] โดยที่แบ่งเว็บคลาวเลอร์เป็น 2 ระดับช่วยกันทำงานคือ

1) คลาวเลอร์ภายใน (Internal Crawler) เป็นคลาวเลอร์ที่ทำหน้าที่เก็บข้อมูลเว็บเพจภายใต้เว็บไซต์ที่ได้รับมอบหมาย โดยจะเริ่มเก็บเว็บเพจจากหน้าโฮมเพจของเว็บไซต์ จนกระทั่งเก็บเว็บเพจได้ครบทั้งเว็บไซต์หรือเก็บได้ครบตามจำนวน
2) คลาวเลอร์ภายนอก (External Crawler) หรือเว็บคลาวเลอร์ในระดับเว็บไซต์ ทำหน้าที่มอบหมายเว็บไซต์ที่ต้องการเก็บข้อมูลส่งไปให้ไซต์คลาวเลอร์ภายใน

2.3 เพจแร็งค์

เพจแร็งค์เป็นอัลกอริทึมที่วิเคราะห์โครงสร้างการเชื่อมโยง (Link Structure Anlysis) ระหว่างเว็บเพจ มีแนวคิดโดยลอกเลียนแบบจากพฤติกรรมมนุษย์ในเรื่องของการให้คะแนนเสียง (Vote) หรือแนะนำบุคคล (Recommendation) กล่าวคือ บุคคลใดได้รับคะแนนเสียงมาก หรือบุคคลได้รับการแนะนำจากบุคคลที่สำคัญทางสังคมมากๆ ย่อมแสดงว่าบุคคลนั้นมีความสำคัญมาก ด้วยเหตุผลในทำนองเดียวกัน เว็บเพจที่มีการเชื่อมโยงมาจาก หรือถูกชี้ (In-Link) โดยเว็บเพจอื่นๆ เป็นจำนวนมาก ย่อมบ่งบอกให้เห็นว่าเว็บเพจนั้นมีความสำคัญมากเช่นกัน

ในการคำนวณเพจแร็งค์ จะอาศัยการคำนวณแบบวนซ้ำ (Iterative Computation) ที่เรียกว่า วิธีการพาวเวอร์ (Power Method) คะแนนความสำคัญของเว็บเพจ หรือที่เรียกว่า ค่าเพจแร็งค์ (Page Rank Score) นั้นเป็นค่าคะแนนหนึ่งที่ระบบสืบค้นข้อมูลกูเกิ้ล ใช้ในการจัดเรียงลำดับเว็บผลลัพธ์ของการค้นคืน ซึ่งในทางทฤษฎีค่าเพจแร็งค์ของเว็บเพจหนึ่งๆ คำนวณได้จากค่าเพจแร็งค์ของเว็บเพจที่มีเส้นการเชื่อมโยงมาหา กล่าวคือ สำหรับเส้นการเชื่อมโยงจากเว็บเพจ u ใดๆ ไปยังเว็บเพจ v ใดๆ ($u \rightarrow v$) จะมีความหมายโดยนัยว่า ผู้พัฒนาเว็บเพจ u นั้นได้แนะนำหรือให้ความสำคัญกับเว็บเพจ v ปลายทาง โดยคิดเป็นค่าเพจแร็งค์ $\frac{r(u)}{o(u)}$ หน่วยให้กับเว็บเพจ v เมื่อ $r(u)$ แทนค่าเพจแร็งค์ของเว็บเพจ u และ $o(u)$ แทนจำนวนเส้นการเชื่อมโยงออก (Out-Degree) จากเว็บเพจ u จะได้ว่า

$$r(v) = \sum_{u \in B(v)} \frac{r(u)}{o(u)}$$

นั่นคือค่าเพจแร็งค์ของเว็บเพจ v มีค่าเท่ากับผลรวมของค่าเพจแร็งค์จากทุกๆ เว็บเพจ u ที่เชื่อมโยงไปยังเว็บเพจ v แบ่งตามสัดส่วนซึ่งในอัลกอริทึมเพจแร็งค์มาตรฐาน สัดส่วนนี้ถูกกำหนดโดยเท่าๆ กันตามจำนวน

การเชื่อมโยงออก หรือเท่ากับ $\frac{1}{o(u)}$ ซึ่งหากพิจารณาโครงสร้างการเชื่อมโยงกันของเว็บเพจ หรือเว็บกราฟนั้นมักมีส่วนนที่เชื่อมโยงกันเป็นวงรอบ (Cycle) ดังนั้นในทางปฏิบัติค่าเพจแร็งค์จะถูกคำนวณ และส่งผ่านจากเว็บเพจหนึ่งไปยังอีกเว็บเพจหนึ่งแบบวนซ้ำจนกระทั่งถึงค่าเสถียร (Covergence)

ในลำดับถัดไปจะกล่าวถึงสมการในรูปแบบทั่วไปของการคำนวณเพจแร็งค์ กำหนดให้ P เป็นเมตริกซ์ขนาด $n \times n$ แทนเว็บกราฟ เมื่อ n คือจำนวนเว็บเพจทั้งหมด โดยที่แถวและคอลัมน์ของเมตริกซ์กำหนดเป็นเว็บเพจต้นทาง (Source Web Page) และเว็บเพจปลายทาง (Destination Web Page) ตามลำดับ สำหรับค่าในเมตริกซ์ p_{ij} กำหนดด้วยสัดส่วนตามจำนวนการเชื่อมโยงออกดังนี้

$$p_{ij} = \begin{cases} \frac{1}{o(i)} & \text{กรณีเว็บกราฟมีเส้นเชื่อมโยงจากเว็บเพจ } i \text{ ไปยังเว็บเพจ } j \\ \frac{1}{n} & \text{กรณีเว็บเพจ } i \text{ ไม่มีการเชื่อมโยงออกไปโหนดใดๆบนเว็บกราฟ} \\ 0 & \text{กรณีอื่นๆ} \end{cases}$$

และกำหนดให้แร็งค์เวกเตอร์ $\vec{r}$ เป็นเวกเตอร์แนวคอลัมน์ (Column Vector) ขนาด $n\ x\ 1$ แทนค่าเพจแร็งค์ของทุกๆ เว็บเพจสำหรับในการคำนวณจะใช้วิธีพาวเวอร์กล่าวคือ เป็นการคำนวณแบบวนซ้ำบนการคูณระหว่างเมตริกซ์และเวกเตอร์

$$\vec{r}^{(k+1)} = P^T \vec{r}^{(k)}$$

แร็งค์เวกเตอร์ ณ การคำนวณรอบที่ $k + 1$ มีค่าเท่ากับ ผลคูณระหว่างทรานสโพสเมตริกซ์เว็บกราฟ และแร็งค์เว็กเตอร์จากเดิมรอบที่ k โดยที่กำหนดให้แร็งค์เวกเตอร์เริ่มต้น $\vec{r}^{(0)}$ มีค่าเท่ากับ $\left[\frac{1}{n}\right]_{n\times 1}$ ซึ่งการคำนวณจะหยุดเมื่อ $\vec{r}^{(k+1)}$ ลู่เข้าค่าเสถียร หรือมีผลต่างน้อยกว่า δ ที่กำหนดไว้ค่าหนึ่ง

$$\left\|\vec{r}^{(k+1)} - \vec{r}^{(k)}\right\| < \delta$$

ตามทฤษฎีทางคณิตศาสตร์ การคูณเมตริกซ์และเวกเตอร์แบบวนซ้ำด้วยวิธีการพาวเวอร์นั้น สามารถรับประกันได้ว่าค่าเพจแร็งค์ของทุกๆ เว็บเพจ หรือแร็งค์เวกเตอร์จะลู่เข้าค่าเสถียรเสมอ เมื่อเมตริกซ์ P มีคุณสมบัติอะพิริออดิก (Aperiodic) และเออริดิวซิเบิล (Irreducible) ซึ่งในทางปฏิบัติทำได้โดยสร้างเส้นเชื่อมโยงเสมือนไปยังทุกๆ เว็บเพจอย่างทั่วถึงหรือที่เรียกว่า “การกระโดดสุ่ม” (Random Jump) กล่าวคือเมตริกซ์ P จะถูกแก้ไขเป็น P' ดังนี้

$$P' = \alpha P + (1 - \alpha)\vec{e}\vec{v}$$

เมื่อ $\vec{e}$ เป็นเวกเตอร์แนวคอลัมน์ขนาด $n \times 1$ โดยมีค่าเป็น $[1]_{n\times 1}$ และ α คือ ค่าคงที่สำหรับถ่วงน้ำหนัก (Dampling Factor) ระหว่างเส้นการเชื่อมโยงจริงและเส้นการเชื่อมโยงเสมือนแบบทั่วถึง ซึ่งมักกำหนดให้มีค่าเป็น 0.85 [4] โดยสรุป อัลกอริทึมเพจแร็งค์มีสมการในการคำนวณค่าเพจแร็งค์ให้กับทุกๆ เว็บเพจดังนี้

$$\vec{r}^{(k+1)} = P'^{T}\vec{r}^{(k)}$$

2.4 งานวิจัยที่เกี่ยวข้อง

1) Characteristics of the Web of Spain [1]

การวิเคราะห์คุณลักษณะเว็บสเปนถูกแบ่งออกเป็น 3 ระดับคือ ระดับเว็บเพจ ระดับเว็บไซต์ และระดับโดเมน แต่ละระดับจำแนกคุณลักษณะแต่ละด้านออกมาอธิบายพร้อมแสดงกราฟหรือค่าเชิงสถิติที่ได้อาทิ ระดับเว็บเพจมีคุณลักษณะของ URLs, Page Titles, ภาษา ฯลฯ ระดับเว็บไซต์มีคุณลักษณะของปริมาณเว็บเพจ โครงสร้างลิงก์ระหว่างเว็บไซต์ เป็นต้น

2) Introducing the Portuguese web archive initiative [3]

ทางผู้วิจัยได้นำเสนอวิธีการลดขนาดของข้อมูลจัดเก็บข้อมูลเว็บในรูปแบบ compressed arc file format และวิธีตัดเว็บเพจ รวมถึงเนื้อหาที่ไม่น่าสนใจออกได้แก่ Cascading Style Sheets (CSS) และ JavaScript ซึ่งในงานที่ทำสามารถลดปริมาณข้อมูลที่ต้องจัดเก็บจาก 2.8 TB ลงไปเหลือ 2.0 TB

3) Finding Thai Web Pages in Foreign Web Spaces [5]

ผู้วิจัยได้นำเสนอ Language Specific Web Crawler (LSWC) เพื่อรวบรวมข้อมูลเว็บเพจแบบเจาะจงภาษา โดยเลือกใช้ภาษาไทยในการทดลองและวัดผล ตามสมมติฐานที่ว่า เว็บเพจไทยมักจะชี้ไปยังเว็บเพจไทย ตามคุณสมบัติ locality of outlink จึงมีการจัดเรียงลำดับความสำคัญของลิงก์และการตัดลิงก์ของเว็บเพจรวมถึงเว็บไซต์ ที่ไม่นำไปพิจารณาต่อ เพราะคาดว่าจะไม่เกี่ยวข้องกับเว็บเพจที่เป็นภาษาไทย

ซึ่งการจัดเรียงลำดับความสำคัญของลิงก์จะเรียงตามระยะทางของลิงก์ที่ห่างเว็บเพจไทย จากน้อยไปหามาก และเมื่อนำคลาวเลอร์ที่ใช้อัลกอริทึม Breadth First Search, Soft focused, Hard focused และ LSWC มาเปรียบเทียบกันโดยทดสอบบนชุดข้อมูลทดสอบของพวกเขาที่เตรียมขึ้นมาเองโดยการใช้คลาวเลอร์ดึงข้อมูลจากเว็บต่างๆ เป็นจำนวน 14 ล้านเว็บเพจ พบว่า LSWC มีความเจือปนของเว็บเพจที่ไม่ใช่เว็บเพจภาษาไทยต่ำมากตั้งแต่เริ่มต้นเก็บข้อมูล

3 เครื่องมือที่ใช้ในการทำโครงงาน

3.1 เครื่องเซิร์ฟเวอร์ใช้ในการประมวลผล

3.1.1 ด้านฮาร์ดแวร์

1) หน่วยความจำหลัก 16 GB
2) หน่วยความจำรอง 5.5 TB
3) หน่วยประมวล Intel® Xeon® CPU E5620 ความเร็ว 2.4 GHz
4) การ์ดแลน 1000 Mbpss

3.1.2 ด้านซอฟต์แวร์

1) ระบบปฏิบัติการ CentOS 5.5
2) Java SE Development Kit 7
3) Vim editor

3.2 เครื่องคอมพิวเตอร์พกพาใช้ในการพัฒนาโปรแกรม

3.2.1 ด้านฮาร์ดแวร์

1) หน่วยความจำหลัก 4 GB
2) หน่วยความจำรอง 500 GB
3) หน่วยประมวล Intel® CoreTM i3-3120M CPU ความเร็ว 2.5 GHz
4) การ์ดแลน 100 Mbps

3.2.2 ด้านซอฟต์แวร์

1) ระบบปฏิบัติการ Ubuntu 13.04
2) Java SE Development Kit 7
3) Vim editor
4) NetBeans IDE 7.0.1
5) Google Chrome

3.3 ไลบรารีที่ใช้ในการพัฒนา

1) Gephi 0.82
2) Mallet API
3) JSoup 1.7.2
4) Lucene
5) Lang Detect

4 วิธีการดำเนินโครงงาน

4.1 การเก็บรวบรวมเว็บเพจไทย

การเก็บรวบรวมเว็บเพจไทยจะพัฒนาคลาวเลอร์แบบแยกเก็บตามไซต์ขึ้นมา โดยการเก็บข้อมูลจะเก็บในรูปโครงสร้างของ Arc File Format และมีการเพิ่มเงื่อนไขขึ้นมาจากเดิมคือหากพบว่าเว็บไซต์นั้นไม่ได้มีชื่อโดเมนที่ลงท้ายด้วย .th และไม่ได้มีไอพีแอดเดรสอยู่ในประเทศไทย เราจะทำการตรวจสอบภาษาของเว็บเพจภายในเว็บไซต์นั้นบางส่วน หากพบว่ามีเว็บเพจภาษาไทยอยู่ ก็จะทำการดาวน์โหลดเว็บเพจในเว็บไซต์นั้นมาจนครบจำนวน แต่หากไม่พบเว็บเพจภาษาไทยก็จะไม่ดาวน์โหลดเว็บเพจในเว็บไซต์นั้นต่อ ซึ่งการกำหนดค่าเว็บคลาวเลอร์เป็นดังนี้

1) เลือกยูอาร์แอลต้นทางเป็นเว็บไซต์ภาษาไทยจำนวน 90,000 ลิงก์ จากงานวิจัยของปุณณวัฒน์ [6]
2) กำหนดจำนวนเว็บเพจที่ใช้ตรวจสอบภาษาไทย 5 เว็บเพจ
3) กำหนดจำนวนการดาวน์โหลดสูงสุด 1,000 เว็บเพจต่อ 1 เว็บไซต์

4.2 การวิเคราะห์ตำแหน่งที่ตั้งของเซิร์ฟเวอร์

การวิเคราะห์ตำแหน่งของเซิร์ฟเวอร์เราสามารถทำได้โดยการแปลงชื่อโฮสต์ให้เป็นไอพีแอดเดรสด้วยกระบวนการ DNS Lookup แล้วนำไอพีแอดเดรสที่ได้ไปตรวจสอบตำแหน่งประเทศด้วย GeoIP จะได้ชื่อประเทศที่เป็นตำแหน่งที่ตั้งของเซิร์ฟเวอร์

ในส่วนนี้เราจะทำการตรวจสอบหาความสัมพันธ์ตำแหน่งที่ตั้งของเซิร์ฟเวอร์ที่เป็นเว็บภาษาไทยได้ โดยเราจะได้เป็นข้อมูลสถิติของจำนวนเว็บไซต์ภาษาไทยที่ปรากฏในประเทศอื่นๆ

4.3 การหาโครงสร้างลิงก์การเชื่อมโยงเว็บ

การที่เราจะนำลิงก์ไปวิเคราะห์ หรือคำนวณได้นั้นเราจำเป็นอย่างยิ่งที่จะต้องแปลงลิงก์ให้อยู่ในรูปเชิงตัวเลข ทางผู้วิจัยได้พิจารณาความสัมพันธ์การเชื่อมโยงแบ่งเป็นสองระดับ คือระดับเว็บเพจ และระดับเว็บไซต์ดังนี้

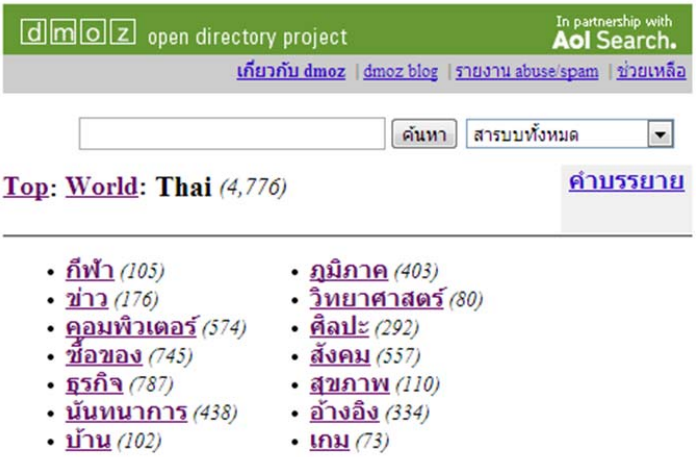
1) ระดับเว็บเพจ กำหนดหมายเลขไอดีเอกสาร (Document ID) ให้กับแต่ละเว็บเพจ แล้วแปลงลิงก์ที่ชี้ไปยังเอกสารต่างๆ ให้เป็นหมายเลขไอดีของเอกสารเหล่านั้น ซึ่งจะได้ความสัมพันธ์โดยโครงสร้างการเชื่อมโยงระหว่าง ไอดีเว็บเพจต้นทาง (S) กับไอดีเว็บเพจปลายทาง (D)
2) ระดับเว็บไซต์ กำหนดหมายเลขไอดีเอกสารให้กับเว็บไซต์ และมองลิงก์ทุกลิงก์ให้เป็นลิงก์ที่ชี้ไปยังเว็บไซต์แทน (ตัดการพิจารณาพาธของลิงก์) แล้วแปลงลิงก์ที่ชี้ไปยังเว็บไซต์ต่างๆ ให้เป็นหมายเลขไอดีของเว็บไซต์นั้นๆ ซึ่งในระดับเว็บไซต์จะมีการถ่วงน้ำหนักของลิงก์ (W) ตามจำนวนเว็บเพจภายในเว็บไซต์ไอดีต้นทาง (S) ที่มีลิงก์ชี้ไปหาเว็บไซต์ไอดีปลายทาง (D)

4.4 การวิเคราะห์ลิงก์

เมื่อเราได้โครงสร้างลิงก์การเชื่อมโยงเว็บมาจากข้อ 4.3 แล้วเราก็จะนำลิงก์เหล่านั้นมาวิเคราะห์เพื่อหาค่าเพจแร็งค์เพื่อดูความสำคัญของเว็บเพจ การกระจายตัวของเพจแร็งค์ ตรวจนับจำนวนลิงก์ชี้เข้า (In-Degree) จำนวนลิงก์ชี้ออก (Out-Degree) มาสร้างกราฟและบันทึกสถิติ

4.5 การจัดประเภทเนื้อหาของเว็บเพจภาษาไทย

ประเภทเนื้อหาของเว็บเพจที่เป็นภาษาไทย เราอิงตามเว็บ dmoz.org ซึ่งเป็นเว็บที่ทำการจัดหมวดหมู่ให้กับเว็บไซต์ทั่วโลก ซึ่งเว็บประเทศไทยถูกจัดไว้ 14 หมวดหมู่ด้วยกัน ดังรูปที่ 2 และเนื่องจากเว็บผิดกฎหมายโดยปกติจะไม่มีการนำมาบรรจุในเว็บ แต่ในการสำรวจเว็บเพจที่เก็บมาพบว่ามีเว็บที่ผิดกฎหมาย อย่างเว็บการพนัน ดังนั้น ทางผู้จัดทำจึงได้เพิ่มประเภทเว็บการพนันขึ้นมาอีก 1 ประเภทรวมเป็น 15 หมวดหมู่



รูปที่ 2 หน้าเว็บ dmoz.org ที่มีการจัดประเภทของเว็บเพจไทย

กระบวนการคัดแยกเว็บเพจ เราจะใช้ MALLET API[9] เป็นไลบรารีเครื่องจักรเรียนรู้ เพื่อคัดแยกหมวดหมู่ของเว็บเพจ โหลดข้อมูลเว็บเพจที่ถูกจัดไว้แล้วในเว็บ dmoz.org มาทำเป็นชุดข้อมูลฝึกสอน โดยที่เราจำเป็นอย่างยิ่งที่จะต้องกรองเนื้อหาออกมาจากเว็บเพจ และตัดคำภาษาไทยให้แยกออกจากกันก่อนที่จะเริ่มฝึกสอน ซึ่งมีขั้นตอนการฝึกสอนดังนี้

1) คัดลอกเว็บเพจบางส่วนที่มีเนื้อหาเกี่ยวกับการพนันที่พบแยกมาเก็บไว้ในแฟ้มแยกประเภท
2) ใช้เว็บคลาวเลอร์ดาวน์โหลดข้อมูลจากหน้าเว็บเพจแต่ละหมวดหมู่ภายใต้ยูอาร์แอล http://www.dmoz.org/World/Thai/ มาแล้วเก็บเอกสารที่ดาวน์โหลดมาได้แยกแฟ้มไว้ตามประเภท
3) ลบแท็กและข้อคิดเห็นออกจากเอกสารเว็บเพจโดยใช้ JSoup [7] ซึ่งเป็นตัววิเคราะห์และจัดการเอกสารที่เป็นโครงสร้าง HTML

4) ตัดคำไทยด้วย LexTo [8] ซึ่งเป็นตัวตัดคำตามพจนานุกรมไทยของ LEXiTRON

5) เริ่มทำการฝึกสอนโดยใช้ MALLET ซึ่งเราสามารถเรียกใช้แบบคอมมานด์ไลน์ในระบบปฏิบัติการลินุกซ์ได้เลยเพราะ MALLET จะเรียนรู้การจัดหมวดหมู่ตามชื่อแฟ้มอยู่แล้ว

เมื่อทำการฝึกสอนเสร็จจนครบ 5 ขั้นตอนแล้วก็จะทดสอบประสิทธิภาพของตัว MALLET หากประสิทธิภาพไม่ดีเท่าที่ควรก็จำเป็นต้องมีการปรับชุดข้อมูลฝึกสอนใหม่ จนเป็นที่ยอมรับได้แล้วจึงนำไปคัดแยกเว็บเพจภาษาไทย

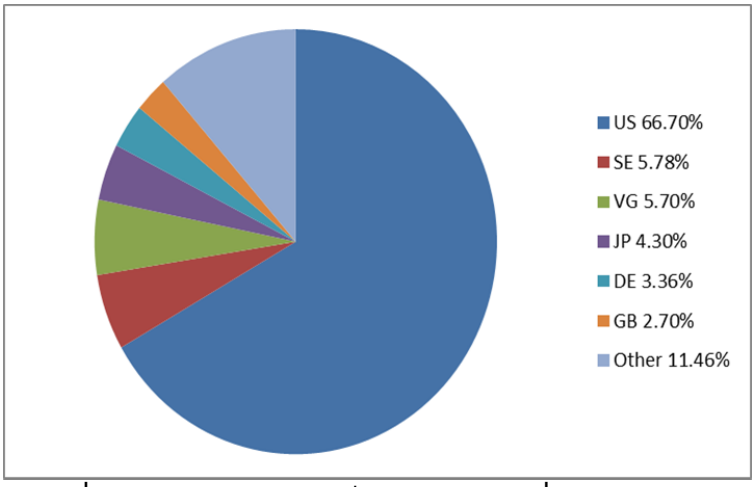
4.6 การวิเคราะห์ความสัมพันธ์ของประเภทเนื้อหาเว็บเพจ

หลังจากที่เราสามารถแยกประเภทของเว็บเพจได้แล้วเราก็สามารถที่จะทราบถึงความนิยมของเว็บไทยว่า ปกติแล้วเนื้อหาเว็บประเภทใดเป็นที่นิยมมากที่สุดสำหรับคนไทย เมื่อดูจากค่าเพจแร็งค์ที่มีสูงอันดับต้นๆ การนำค่ามอดูลาริตี ที่แบ่งเว็บไซต์ออกเป็นชุมชนมาวิเคราะห์ ถึงลักษณะเนื้อหาของเว็บเพจที่อยู่ชุมชนเดียวกัน ทำให้เราทราบถึงความเกี่ยวข้องของประเภทเนื้อหาต่างๆ กล่าวคือเนื้อหาอะไรสัมพันธ์กับอะไรมากที่สุดโดยเทียบจากอัตราส่วนในชุมชน การพิจารณาความสัมพันธ์ระหว่างที่ตั้งของเซิร์ฟเวอร์ กับเนื้อหาอาจทำให้เราได้ค้นพบความรู้อะไรบางอย่าง เช่น เว็บไซต์ผิดกฎหมายมักจะอยู่ในประเทศที่กฎหมายไม่ครอบคลุมในเรื่องนั้น เป็นต้น

5 ผลการดำเนินโครงงานและวิจารณ์

5.1 เว็บไซต์ภาษาไทยที่อยู่ต่างประเทศ

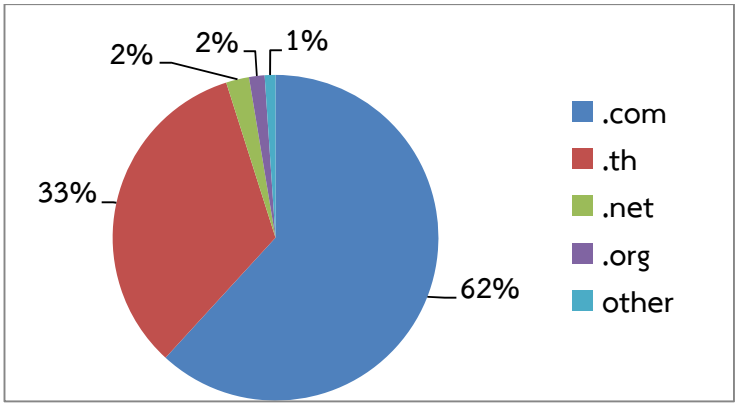
จากเว็บไซต์ภาษาไทยจำนวน 65,732 เว็บไซต์ เป็นเว็บไซต์ที่ตั้งอยู่ในประเทศไทยจำนวน 61,220 เว็บไซต์ (93.14%) และเป็นเว็บไซต์ที่ตั้งอยู่ต่างประเทศ 4,512 เว็บไซต์ (6.86%) เมื่อพิจารณาเฉพาะเว็บไซต์ภาษาไทยที่ตั้งอยู่ต่างประเทศ จากกราฟในรูปที่ 3 พบว่า สหรัฐอเมริกาเป็นประเทศที่มีเว็บไซต์ภาษาไทยปรากฏมากที่สุดในต่างประเทศจำนวน 3,009 เว็บไซต์ (66.70%) รองลงมาคือสวีเดน (5.78%) หมู่เกาะบริติชเวอร์จิน (5.70%) ญี่ปุ่น (4.30%) เยอรมัน (3.36%) และอังกฤษ (2.70%) ตามลำดับ



รูปที่ 3 กราฟแสดงสัดส่วนเว็บไซต์ภาษาไทยที่อยู่ต่างประเทศ

5.2 สัดส่วน TLD ของเว็บไซต์ภาษาไทย

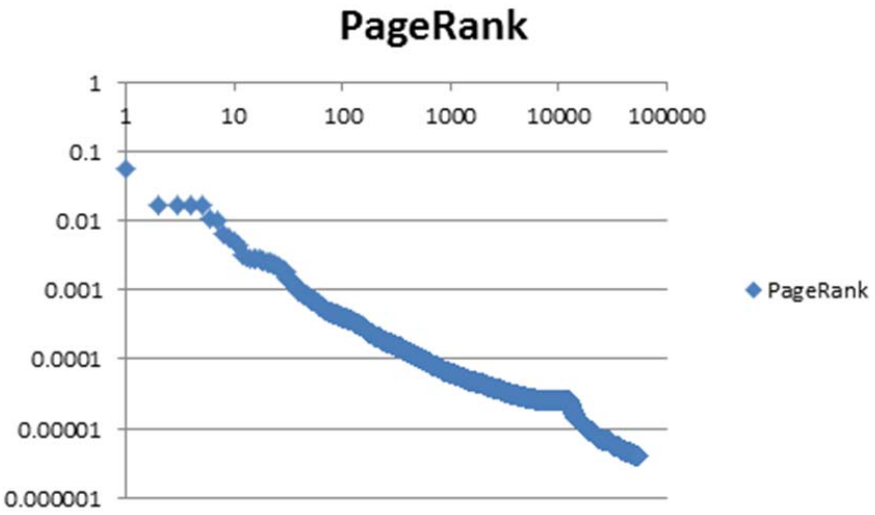
เมื่อเทียบสัดส่วน Top level domain ของเว็บไซต์ภาษาไทยตามรูปที่ 4 จะพบว่าเว็บไทยส่วนใหญ่จะใช้โดเมนเนม .com เป็นจำนวน 40,628 เว็บไซต์ รองลงมาก็จะเป็น .th เป็นจำนวน 21,881 เว็บไซต์ .net เป็นจำนวน 1,512 เว็บไซต์ .org เป็นจำนวน 1,011 เว็บไซต์ ซึ่งแปลว่า .com และ .th เป็นโดเมนที่นิยมจดกันสูงสุดในเว็บภาษาไทย



รูปที่ 4 กราฟแสดงสัดส่วน TLD ของเว็บไซต์ภาษาไทย

5.3 การกระจายตัวเพจแร็งค์

เมื่อพิจารณาค่าการกระจายตัวเพจแร็งค์ของเว็บไทยแล้วพบว่ามีการกระจายตัวในลักษณะของ power law กล่าวคือเว็บที่มีค่าเพจแร็งค์สูงๆ จะมีจำนวนน้อยมาก ส่วนเว็บที่มีค่าเพจแร็งค์ต่ำๆ จะมีจำนวนมากขึ้นมามากขึ้นตามลำดับ



รูปที่ 5 การกระจายตัวเพจแร็งค์ของเว็บไทย

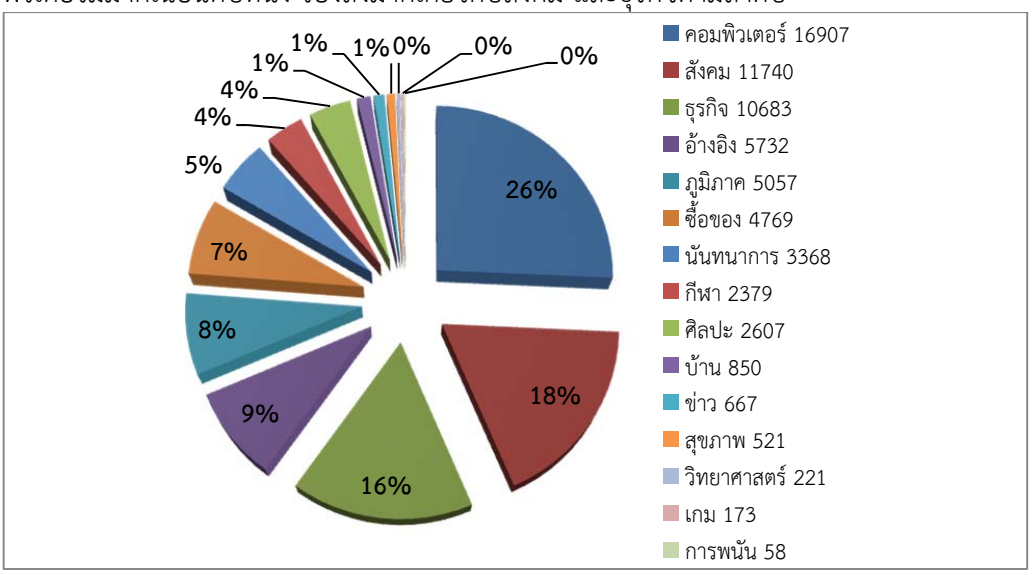
5.4 การคัดแยกหมวดหมู่เว็บไซต์ภาษาไทย

การจัดหมวดหมู่เว็บไซต์ภาษาไทยโดยใช้ Mallet API เมื่อใช้ 10-fold cross validation ได้ผลดังตารางที่ 1 พบว่าเราสามารถจัดหมวดหมู่เว็บไซต์ได้โดยมีค่าเฉลี่ยประสิทธิภาพโดยรวม (F-Measure) อยู่ที่ 0.837

ตารางที่ 1 ผลการทดสอบ 10-fold cross validataion กับข้อมูลฝึกสอนการจำแนกเว็บไทย

หมวดหมู่	Precision	Recall	F-Measure
สุขภาพ	0.863874	0.921788	0.891892
ข่าว	0.750000	0.716667	0.732955
วิทยาศาสตร์	0.708075	0.870229	0.780822
กีฬา	0.804878	0.891892	0.846154
การพนัน	0.983193	0.983193	0.983193
สังคม	0.834805	0.849807	0.842239
อ้างอิง	0.875000	0.880795	0.877888
ภูมิภาค	0.702899	0.659864	0.680702
ศิลปะ	0.884793	0.874715	0.879725
ธุรกิจ	0.813754	0.783088	0.798126
นันทนาการ	0.926635	0.888379	0.907104
คอมพิวเตอร์	0.875429	0.892774	0.884016
บ้าน	0.690476	0.874372	0.771619
ซื้อของ	0.945706	0.880515	0.911947
เกม	0.664384	0.906542	0.766798
เฉลี่ย	0.821593	0.858308	0.837012

จากคัดแยกหมวดหมู่เว็บไซต์ภาษาไทยมีสัดส่วนดังรูปที่ 6 ซึ่งพบว่ามีปริมาณเนื้อหาที่พบเกี่ยวกับคอมพิวเตอร์มีมากเนอันดับหนึ่ง รองลงมาก็เกี่ยวกับสังคม และธุรกิจตามลำดับ



รูปที่ 6 สัดส่วนเว็บไซต์ภาษาไทยแยกตามเนื้อหา

6 สรุปผลการดำเนินงานและข้อเสนอแนะ

6.1 ข้อสรุป

โครงงานนี้จัดทำขึ้นเพื่อหาคุณลักษณะเว็บไซต์ของประเทศไทย ที่นำไปใช้ในการพัฒนาเสิร์จเอ็นจินได้ อาทิเช่น ในโครงงานนี้เราสามารถจัดประเภทเนื้อหาของแต่ละเว็บเพจได้ เราก็เราสามารถเอาคุณสมบัตินี้ มาทำเป็นตัวเลือกในการค้นหาเพื่อใช้ในการกรองคำค้นที่กำกวมได้ ซึ่งเสิร์จเอ็นจินทั่วไปเรามักจะไม่สามารถเลือกประเภทเนื้อหาในการค้นหา

6.2 ข้อเสนอแนะและงานต่อเนื่อง

เนื่องจากการเก็บข้อมูลที่นำมาใช้ในการวิเคราะห์เว็บไทยนั้นเป็นเพียงข้อมูลในช่วงเวลาหนึ่งเท่านั้น แต่โครงงานในชิ้นนี้ได้มีการออกแบบให้สามารถนำมาวิเคราะห์เว็บไทยเพิ่มเติมในภายหลังได้อีก ซึ่งเราความจริงแล้วเราควรศึกษาคุณลักษณะการเปลี่ยนแปลงของเว็บเพจไทยร่วมด้วย เพื่อพิจารณาแนวโน้มต่างๆ ที่เกิดขึ้นเกี่ยวกับเว็บไทย รวมถึงการนำความรู้ที่ได้ไปใช้ในการพัฒนาเว็บคลาวเลอร์ และเสิร์จเอ็นจินไทยในอนาคต

6.3 ปัญหาและอุปสรรคที่พบ

การวิเคราะห์ลิงก์การเชื่อมโยงต้องใช้หน่วยความจำหลักเยอะและใช้เวลานานในการวิเคราะห์ต้องใช้เวลารอหลายวัน และการเพิ่มข้อมูลเข้าไปวิเคราะห์เพิ่มเติมเพียงบางส่วน ต้องเริ่มประมวลผลใหม่เกือบทั้งหมด

7 บรรณานุกรม

[1] R. Baeza-Yates, C. Castillo, and V. Lopez. Characteristics of the web of spain. International Journal of Scientometrics, Informetrics and Bibliometrics, 9(1), 2005.

[2] M. Ester, H. Kriegel and M. Schubert, “Accurate and efficient crawling for relevant website”, In Proc. VLDB ’04 Proceedings of the Thirtieth international conference on Very large data base – Volume 30, 2004.

[3] D. Gomes, A. Nogueira, J. Miranda, and M. Costa, “Introducing the Portuguese web archive initiative”, in 8th International Web Archiving Workshop (IWAW08), Aarhus, Denmark, September 2008.

[4] T. H. Haveliwala, “Efficient Computation of PageRank”, Stanford University, Technical Report, 1999.

[5] K. Somboonviwat, T. Tamura, and M. Kitsuregawa, “Finding Thai web pages in foreign web spaces”, In Proc. of the 22nd Inter. Conference on Data Engineering Workshops, 2006.

[6] P. Tadapak, T. Suebchua, and A. Rungsawang. A machine learning based language specific web site crawler. In Proceedings of the 13th International Conference on Network-Based Information Systems, pages 155–161, 2010.

[7] JSoup HTML Parser, http://jsoup.org/., Accessed on August 2013.

[8] LexTo, http://www.sansarn.com/lexto/., Accessed on September 2013.

[9] MAchine Learning for LanguagE Toolkit (MALLET), http://mallet.cs.umass.edu/., Accessed on September 2013.

8 ภาคผนวก

ในส่วนนี้จะกล่าวถึงวิธีการใช้งานโปรแกรมวิเคราะห์คุณลักษณะเว็บไทย ซึ่งมีขั้นตอน 3 ขั้นตอนหลักๆ ดังนี้

1) เก็บข้อมูล (Crawling) ทำหน้าที่เก็บข้อมูลเว็บไซต์ที่ได้รับมอบหมาย

2) สกัดฟีเจอร์ (Feature Extraction) ทำหน้าที่สกัดคุณลักษณะฟีเจอร์จากเว็บไซต์ที่เก็บได้เพื่อเตรียมใช้สำหรับการวิเคราะห์ลิงก์และเนื้อหาของเว็บไทย

3) การวิเคราะห์ (Analysis) ทำหน้าที่คำนวณค่าเชิงสถิติพื้นฐานของเว็บไทย รวมถึงการจัดหมวดหมู่ของเนื้อหา

ซึ่งแต่ละขั้นตอนมีการเรียกคำสั่งที่ต่างกัน ซึ่งมีรายละเอียดวิธีการใช้และเรียกคำสั่งแต่ละขั้นตอนดังนี้

8.1 การรันขั้นตอน Crawling

8.1.1 การนำเข้า Seed Site

ในการนำเข้า Seed Site จะใช้คำสั่ง

```
java -cp "WebAnalysis.jar;lib/*" Crawler.Import <Seed File>
```

- <Seed File> คือที่อยู่ของไฟล์ Seed Site โดยมีรูปแบบคือมีชื่อเว็บไซต์บรรทัดละหนึ่งเว็บไซต์ โดยไม่ต้องมีโปรโตคอลนำหน้า (ตัด http:// ทิ้ง)
- ตัวอย่างรูปแบบของ Seed File

```
www.ku.ac.th
www.cpe.ku.ac.th
www.sanook.com
www.kapook.com
```

- หลังการนำเข้า Seed ไฟล์ข้อมูลเว็บไซต์จะถูกบันทึกลงในฐานข้อมูลโปรแกรม

8.1.2 คำสั่งเรียกใช้ Crawler ในการเก็บข้อมูล

```
java -cp "WebAnalysis.jar;lib/*" Crawler.Main [<Site Limit>]
```

- <Site Limit> คือจำนวนเว็บไซต์ที่จะคลาวทั้งหมดที่ต้องการ โดยปกติถ้าไม่กำหนด เว็บคลาวเลอร์จะเก็บมาแค่เฉพาะ Seed Site กล่าวคือ Site Limit จะมีค่าเท่ากับ จำนวน Seed Site โดยปริยาย
- โดยข้อมูลที่ได้จากการดาวน์โหลดจะถูกบันทึกในรูปแบบ arc ไฟล์ฟอร์แมท และบีบอัดด้วย gzip มีชื่อไฟล์เป็น crawl-<ชื่อเว็บไซต์>.arc.gz ไว้ในโฟลเดอร์ที่ชื่อว่า Crawler โดย 1 ไฟล์จะเก็บข้อมูล 1 เว็บไซต์

8.2 การรันขั้นตอน Feature Extraction

หลังจากที่การรันคำสั่งในขั้นตอน Crawling เสร็จสมบูรณ์แล้ว เราก็จะเรียกคำสั่งที่ใช้ในการสกัดฟีเจอร์ก่อนไปวิเคราะห์โดยใช้คำสั่ง

java -cp "WebAnalysis.jar;lib/*" Extraction.Main [all|link|content]

- [all|link|content] เป็นตัวเลือกการสกัดฟีเจอร์ แต่ถ้าไม่กำหนดตัวเลือกจะมีค่าเป็น all โดยปริยาย
  - all สกัดฟีเจอร์ทั้งหมด
  - link สกัดฟีเจอร์สำหรับนำไปใช้วิเคราะห์ลิงก์และการเชื่อมโยง
  - content สกัดฟีเจอร์สำหรับนำไปใช้วิเคราะห์เนื้อหา
- การสกัดฟีเจอร์ จะสกัดจากข้อมูลเว็บ .arc.gz จากโฟลเดอร์ Crawler ที่ได้จากข้อ 8.1.2
- ผลลัพธ์ที่ได้ทุกอย่างจะถูกเก็บไว้ที่โฟลเดอร์ที่ชื่อว่า Feature
- ผลลัพธ์ที่ได้จากการสกัดลิงก์ จะเป็นลักษณะของ csv ไฟล์แต่ละบรรทัดแสดงถึงการเชื่อมโยงออกจากโหนดต้นทางคั่นด้วยเซมิโคลอน ; ตามด้วยโหนดปลายทางที่ 1 ; โหนดปลายทางที่ 2 ; ไปจนครบทุกโหนด ถูกบันทึกเป็นไฟล์ชื่อว่า LinkPage.csv ในโฟลเดอร์ Graph
- ผลลัพธ์ที่ได้จากการสกัดฟีเจอร์เนื้อหา จะเป็นลักษณะของ arc ไฟล์ถูกบีบอัดด้วย gzip โดยจะสกัดเฉพาะคำที่ปรากฏในหน้าเว็บเพจ กล่าวคือจะตัดแท็ก HTML ทิ้งและจะทำการตัดคำให้แต่ละคำถูกคั่นด้วยช่องว่าง ซึ่งผลลัพธ์จะมีชื่อไฟล์คงเดิมแต่เปลี่ยนไปอยู่ในโฟลเดอร์ Lex

8.3 การรันขั้นตอน Analysis

หลังจากที่การรันคำสั่งในขั้นตอน Crawling เสร็จสมบูรณ์แล้ว เราก็จะเรียกคำสั่งที่ใช้ในการสกัดฟีเจอร์ก่อนไปวิเคราะห์โดยใช้คำสั่ง

java -cp "WebAnalysis.jar;lib/*" Analysis.Main [option]

- [option] เป็นตัวเลือกการวิเคราะห์ แต่ถ้าไม่กำหนดตัวเลือกจะมีค่าเป็น all โดยปริยาย
  - all วิเคราะห์ทั้งการเชื่อมโยงลิงก์ และเนื้อหา
  - pagerank หาค่าเพจแรงค์
  - iolink หาค่า in-degree และ out-degree ของเว็บเพจ
    - inlink หาค่า in-degree ของเว็บเพจ
    - outlink หาค่า out-degree ของเว็บเพจ
  - content วิเคราะห์จัดประเภทเนื้อหา

ประวัตินิสิต

1 ชื่อ-นามสกุล นายวิวัฒน์ นิ่มสนิทนันท์ เลขประจำตัวนิสิต 5310504001
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
ที่อยู่ปัจจุบัน 5 หมู่ 1 ตำบลแวง อำเภอโพนทอง จังหวัดร้อยเอ็ด 45110
โทรศัพท์ ที่บ้าน 0 4353 4135 โทรศัพท์เคลื่อนที่ 08 0442 0858
E-mail wiwat@octknight.com
ระดับการศึกษา:

คุณวุฒิการศึกษา	จากโรงเรียน/สถาบัน	ปีการศึกษาที่จบ
มัธยมศึกษาตอนปลาย	โรงเรียนโพนทองพัฒนาวิทยา	2552
มัธยมศึกษาตอนต้น	โรงเรียนโพนทองพัฒนาวิทยา	2549