Statistical Study of Thai Web Structure

Piyapong Seibudta, Wiwat Nimsanitnan, Arnon Rungsawang, Bundit Manaskasemsak

Massive Information and Knowledge Engineering Laboratory, Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok 10900, Thailand

E-mail: piya\_tt@hotmail.com, wiwat@octknight.com, arnon@mikelab.net, un@mikelab.net

Abstract

Website on the Internet is growing rapidly and very large. To find information on the Internet requires a tool called Search engine to help to ease and precision. Of the importance of Search engine brings with it a need to determine the structure. Thailand and other features of the web for information on Search engine developed in this research is the study of Thai web. The results of preliminary experiments showed that the statistics can be used to help develop a better Search engine .

**Keywords:**  Web structure, web graph, Search engine

# 1. Introduction

 ปัจจุบันอินเทอร์เน็ต มีความสำคัญต่อชีวิตประจำวันของคนเรา หลายๆ ด้าน ทั้งการศึกษา พาณิชย์ ธุรกรรม วรรณกรรม และอื่นๆ เว็บไซต์ต่างๆ ก็ถูกสร้างขึ้นอย่างมากมายมหาศาล จากการศึกษาของ Netcraft Web Server Survey [1] ประมาณการว่า ในเดือนธันวาคม ค.ศ.2012 มีประมาณ 633.7 ล้านเว็บไซต์ ด้วยขนาดที่ใหญ่มากและการเปลี่ยนแปลงที่มีอยู่ตลอดเวลาทำให้การศึกษาโครงสร้างของทำได้ยาก Broder และคณะ [2] ได้นำเสนอแม่แบบของโครงสร้างเว็บที่เรียกว่า “โครงสร้างรูปโบว์ “(bow-tie structure) ขึ้นมา สำหรับการวิเคราะห์โครงสร้างการเชื่อมโยงของเว็บ ซึ่งช่วยให้เข้าใจภาพรวมโครงสร้างการเชื่อมโยงและค่าสถิติที่สำคัญต่างๆ อีกด้วย  
 ผลวิจัยการศึกษาโครงสร้างเว็บไซต์ของหลายๆ ประเทศแสดงให้เห็นว่าโครงสร้างมีลักษณะที่คล้ายกัน โดยประโยชน์ที่ได้รับจากการศึกษาโครงสร้างเว็บไซต์นั้นมีมากมาย อาทิเช่น การศึกษาชุมชนเว็บออนไลน์ภายในประเทศ การพัฒนาเครื่องมือสำหรับการวิจัยทางสังคมออนไลน์ และวิเคราะห์การตลาดออนไลน์ และที่สำคัญคือเพื่อเป็นการช่วยในการพัฒนา Search engine ให้ดียิ่งขึ้น  
 จากข้อมูลมากมายมหาศาลที่อยู่บนอินเทอร์เน็ต การที่จะค้นหาข้อมูลจำนวนมากมายเช่นนี้เราไม่อาจจะคลิกเพื่อค้นหาข้อมูลพบได้ง่ายๆ จำเป็นจะต้องอาศัยการค้นหาข้อมูลด้วยเครื่องมือค้นหาที่เรียกว่า Search engine เข้ามาช่วยเพื่อความสะดวกรวดเร็ว และที่สำคัญคือความแม่นยำในการจัดเรียงข้อมูลให้ตรงกับความต้องการของผู้สืบค้น จากความสำคัญของ Search engine ดังกล่าวทำให้คณะทำงานในห้องปฏิบัติการวิจัยวิศวกรรมข้อมูลและฐานความรู้ขนาดใหญ่ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์มหาวิทยาลัย เกษตรศาสตร์ (Mike lab) มีความสนใจที่จะทำ Search engine สำหรับเว็บประเทศไทยขึ้น   
 โดยในงานวิจัยนี้จะเป็นการศึกษาเว็บไซต์ไทย โดยจะดำเนินการศึกษาโครงสร้างและข้อมูลสถิติที่สำคัญ เพื่อเป็นข้อมูลพื้นฐานในการพัฒนา Search engine สำหรับภาษาไทยต่อไป

**2. Thai Web Graph and Thai Web Dataset**

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

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

1. เลือกยูอาร์แอลต้นทางเป็นเว็บไซต์ภาษาไทยจำนวน 90,000 ลิงก์ จากงานวิจัยของปุณณวัฒน์ และคณะ [3]

2. กำหนดจำนวนเว็บเพจที่ใช้ตรวจสอบว่าเป็นเว็บภาษาไทย 5 เว็บเพจต่อ 1 ไซต์

3. กำหนดจำนวนการดาวน์โหลดสูงสุด 1,000 เว็บเพจต่อ 1 เว็บไซต์

จากข้อมูลที่จัดเก็บได้จากเงื่อนไขที่กำหนดข้างต้น เราได้ดาวน์โหลดเว็บภาษาไทยรวมทั้งสิ้น ….เว็บเพจ

**2.1 การวิเคราะห์ตำแหน่งที่ตั้งของเซิร์ฟเวอร์**

การวิเคราะห์ตำแหน่งของเซิร์ฟเวอร์เราสามารถทำได้โดยการแปลงชื่อโฮสต์ให้เป็นไอพีแอดเดรสด้วยกระบวนการ DNS Lookup แล้วนำไอพีแอดเดรสที่ได้ไปตรวจสอบตำแหน่งประเทศด้วย GeoIP [4] จะได้ชื่อประเทศที่เป็นตำแหน่งที่ตั้งของเซิร์ฟเวอร์ในส่วนนี้เราจะทำการตรวจสอบหาความสัมพันธ์ตำแหน่งที่ตั้งของเซิร์ฟเวอร์ที่เป็นเว็บภาษาไทยได้ โดยเราจะได้เป็นข้อมูลสถิติของจำนวนเว็บไซต์ภาษาไทยที่ปรากฎในประเทศอื่นๆ

**2.2 การหาโครงสร้างลิงก์การเชื่อมโยงเว็บ**

การที่เราจะนำลิงก์ไปวิเคราะห์ หรือคำนวณได้นั้นเราจำเป็นอย่างยิ่งที่จะต้องแปลงลิงก์ให้อยู่ในรูปเชิงตัวเลข ทางผู้วิจัยได้พิจารณาความสัมพันธ์การเชื่อมโยงแบ่งเป็นสองระดับ คือระดับเว็บเพจ และระดับเว็บไซต์

ดังนี้

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)

**2.2.1 การวิเคราะห์ลิงก์**

เมื่อเราได้โครงสร้างลิงก์การเชื่อมโยงเว็บมาจากข้อ 2.2 แล้วเราก็จะนำลิงก์เหล่านั้นมาวิเคราะห์เพื่อหาค่าเพจแร็งค์เพื่อดูความสำคัญของเว็บเพจ การกระจายตัวของเพจแร็งค์ ตรวจนับจำนวนลิงก์ชี้เข้า (In-Degree) จำนวนลิงก์ชี้ออก (Out-Degree) มาสร้างกราฟและบันทึกสถิติ

**2.3 การจัดประเภทเนื้อหาของเว็บเพจภาษาไทย**

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

****

รูปที่ 1 การจัดประเภทเนื้อหาของเว็บประเทศไทยโดย dmoz.org

กระบวนการคัดแยกเว็บเพจ เราจะใช้ MALLET API [5] เป็นไลบรารีเครื่องจักรเรียนรู้ เพื่อคัดแยกหมวดหมู่ของเว็บเพจ โหลดข้อมูลเว็บเพจที่ถูกจัดไว้แล้วในเว็บ dmoz.org มาทำเป็นชุดข้อมูลฝึกสอน โดยที่เรา

จำเป็นอย่างยิ่งที่จะต้องกรองเนื้อหาออกมาจากเว็บเพจ และตัดคำภาษาไทยให้แยกออกจากกันก่อนที่จะเริ่มฝึกสอน ซึ่งมีขั้นตอนการฝึกสอนดังนี้

1. คัดลอกเว็บเพจบางส่วนที่มีเนื้อหาเกี่ยวกับการพนันที่พบแยกมาเก็บไว้ในแฟ้มแยกประเภท

2. ใช้เว็บคลาวเลอร์ดาวน์โหลดข้อมูลจากหน้าเว็บเพจแต่ละหมวดหมู่ภายใต้ยูอาร์แอล http://www.dmoz.org/World/Thai/ มาแล้วเก็บเอกสารที่ดาวน์โหลดมาได้แยกแฟ้มไว้ตามประเภท

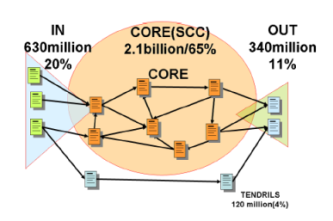
3. ลบแท็กและข้อคิดเห็นออกจากเอกสารเว็บเพจโดยใช้ JSoup [6] ซึ่งเป็นตัววิเคราะห์และจัดการเอกสารที่เป็นโครงสร้าง HTML

4. ตัดคำไทยด้วย LexTo [7] ซึ่งเป็นตัวตัดคำตามพจนานุกรมไทยของ LEXiTRON

5. เริ่มทำการฝึกสอนโดยใช้ MALLET [5] ซึ่งเราสามารถเรียกใช้แบบคอมมานด์ไลน์ในระบบปฏิบัติการลินุกซ์ได้เลยเพราะ MALLET จะเรียนรู้การจัดหมวดหมู่ตามชื่อแฟ้มอยู่แล้ว เมื่อทำการฝึกสอนเสร็จจนครบ 5 ขั้นตอนแล้วก็จะทดสอบประสิทธิภาพของตัว MALLET หากประสิทธิภาพไม่ดีเท่าที่ควรก็จำเป็นต้องมีการปรับชุดข้อมูลฝึกสอนใหม่ จนเป็นที่ยอมรับได้แล้วจึงนำไปคัดแยกเว็บเพจภาษาไทย

1. **Link Structure Statistical**

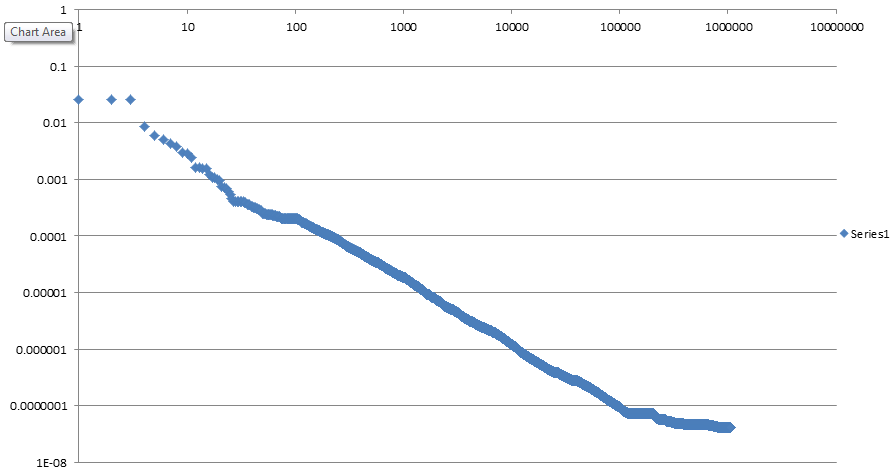
**3.1 โครงสร้างรูปโบว์ (Bow-tie structure)** เมื่อพิจารณาการเชื่อมโยงของเว็บภาษาไทยพบว่ามีโครงสร้างเป็นรูปโบว์ มีลักษณะแบบไม่สมมาตร คือมีส่วนเชื่อมเข้าที่ใหญ่กว่าส่วนเชื่อมออกมาก ดังรูปที่ 2

****

รูปที่ 2 โครงสร้างรูปโบว์ของเว็บภาษาไทย

**3.2 การกระจายตัวเพจแร็งค์**

เมื่อพิจารณาค่าการกระจายตัวเพจแร็งค์ของเว็บไทยแล้วพบว่ามีการกระจายตัวในลักษณะของ power law กล่าวคือเว็บที่มีค่าเพจแร็งค์สูงๆ จะมีจำนวนน้อยมาก ส่วนเว็บที่มีค่าเพจแร็งค์ต่ำๆ จะมีจำนวนมากขึ้นมามากขึ้นตามลำดับดังรูปที่ 3

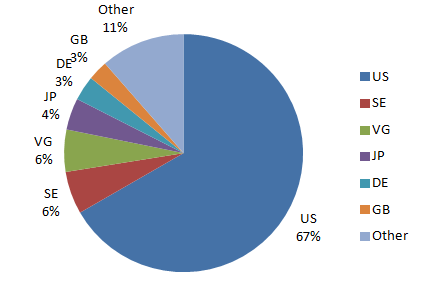


รูปที่ 3 การกระจายตัวเพจแร็งค์ของเว็บไทย

1. **Content Statistical**

**4.1 เว็บไซต์ภาษาไทยที่อยู่ต่างประเทศ**

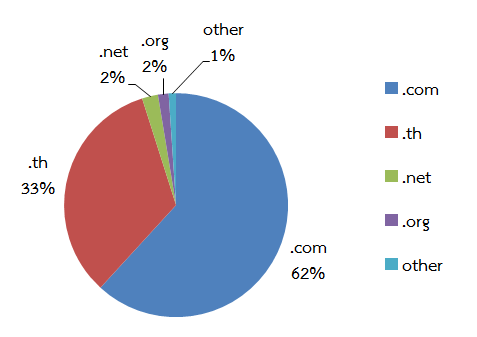
จากเว็บไซต์ภาษาไทยจำนวน 65,732 เว็บไซต์ เป็นเว็บไซต์ที่ตั้งอยู่ในประเทศไทยจำนวน 61,220 เว็บไซต์ (93.14%) และเป็นเว็บไซต์ที่ตั้งอยู่ต่างประเทศ 4,512 เว็บไซต์ (6.86%) เมื่อพิจารณาเฉพาะเว็บไซต์ภาษาไทยที่ตั้งอยู่ต่างประเทศ จากกราฟในรูปที่ 3 พบว่า สหรัฐอเมริกาเป็นประเทศที่มีเว็บไซต์ภาษาไทยปรากฏมากที่สุดในต่างประเทศจำนวน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

****

รูปที่ 4 กราฟแสดงสัดส่วนเว็บไซต์ภาษาไทยที่อยู่ต่างประเทศ

**4.2 สัดส่วน TLD ของเว็บไซต์ภาษาไทย**

เมื่อเทียบสัดส่วน Top level domain ของเว็บไซต์ภาษาไทยตามรูปที่ 2 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

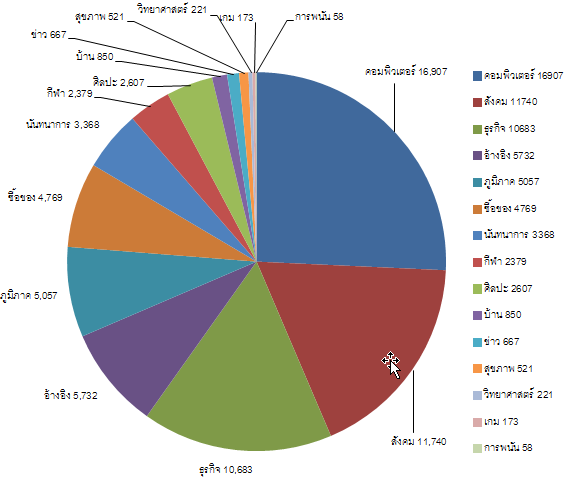


รูปที่ 5 กราฟแสดงสัดส่วนTLD ของเว็บไซต์ภาษาไทย

**4.3 การคัดแยกหมวดหมู่เว็บไซต์ภาษาไทย**

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





รูปที่ 6 สัดส่วนเว็บไซต์ภาษาไทยแยกตามเนื้อหา

จากคัดแยกหมวดหมู่เว็บไซต์ภาษาไทยดังรูปที่ 6 พบว่ามีปริมาณเนื้อหาที่พบเกี่ยวกับคอมพิวเตอร์มีมากเป็นอันดับหนึ่ง รองลงมาก็เกี่ยวกับสังคม และธุรกิจตามลำดับ

**5. Conclusion**

ในงานวิจัยนี้จัดทำเพื่อหาคุณลักษณะเว็บไซต์ของประเทศไทย ที่นำไปใช้ในการพัฒนาเสิร์จเอ็นจินได้ อาทิเช่นในงานวิจัยนี้เราสามารถจัดประเภทเนื้อหาของแต่ละเว็บเพจได้ เราก็เราสามารถเอาคุณสมบัตินี้ มาทำเป็นตัวเลือกในการค้นหาเพื่อใช้ในการกรองคำค้นที่กำกวมได้ ซึ่งเสิร์จเอ็นจินทั่วไปเรามักจะไม่สามารถเลือกประเภทเนื้อหาในการค้นหา

# References

[1]Netcraft: Web Server Survey (December 2012),

<http://news.netcraft.com/archives/2012/12/04/december-2012-web-server-survey.html>

[2] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., State, R., Tomkins, A.,Wiener,J.2000. Graph structure in the web. In Proc. of the 9th World Wide Web Conf.

[3] P. Tadapak,  T. Suebchua and A. Rungsawang. 2010. A Machine Learning Based Language Specific Web Site Crawler, In Proc. of the 13th [Network-Based Information Systems (NBiS)](http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5634971).

[4] GeoIP, <http://en.wikipedia.org/wiki/Geolocation>

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

September 2013.

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

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