รายงานการวิจัยฉบับสมบูรณ์

การครอลเว็บ 2.0 แอปพลิเคชั่นที่มีประสิทธิภาพ
Efficient Crawling of Web 2.0 Applications

ดร. กุลวดี สมบูรณ์วิวัฒน์

ได้รับทุนสนับสนุนงานวิจัยจากงบประมาณเงินรายได้ ประจำปีงบประมาณ 2555
วิทยาลัยนานาชาติ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อโครงการ (ภาษาไทย) การครอลเว็บ 2.0 แอปพลิเคชั่นที่มีประสิทธิภาพ
แหล่งเงิน (ระบุแหล่งทุน) งบประมาณเงินรายได้ วิทยาลัยนานาชาติ
ประจำปีงบประมาณ 2555 จำนวนเงินที่ได้รับการสนับสนุน 50,000 บาท
ระยะเวลาทำการวิจัย 1 ปี ตั้งแต่ 1 ตุลาคม 2554 ถึง 30 กันยายน 2555
ชื่อ-สกุล หัวหน้าโครงการ และผู้ร่วมโครงการวิจัย พร้อมระบุ หน่วยงานต้นสังกัด
ดร. กุลวดี สมบูรณ์วิวัฒน์ หัวหน้าโครงการ วิทยาลัยนานาชาติ (kskulwad@kmitl.ac.th)

บทคัดย่อ

ในปัจจุบัน เวิร์ลไวด์เว็บได้กลายเป็นแพลตฟอร์มของสังคมที่ซึ่งผู้ใช้งานเว็บสามารถสร้างเนื้อหาและแบ่งปันข้อมูล ความคิดเห็นและความคิดสร้างสรรค์ต่างๆ ของตนได้ เว็บ 2.0 เป็นคำศัพท์ที่ได้รับการบัญญัติขึ้นในช่วงปีคริสตศักราช 1999 เพื่อใช้เรียกเทคโนโลยีเว็บยุคใหม่ซึ่งก้าวล้ำกว่าเทคโนโลยีการสร้างเว็บเพจแบบสแตติกในยุคแรก ความแพร่หลายของเทคโนโลยีเว็บ 2.0 ได้นำไปสู่ปัญหาที่เรียกว่า information abundance ซึ่งหมายถึงการเพิ่มขึ้นอย่างรวดเร็วของปริมาณข้อมูลซึ่งมีขนาดมากเกินกว่าความสามารถของมนุษย์ที่จะซึมซับและนำไปใช้ประโยชน์ได้

งานวิจัยนี้ต้องการตอบโจทย์ของปัญหา information abundance ที่เกิดขึ้นกับระบบการครอลเว็บ (web crawling system) โดยแนวคิดหลักในการแก้ไขปัญหาดังกล่าวอยู่ที่การประยุกต์ใช้เทคนิคการทำเว็บคอมมูนิตี้ไมนิ่งเพื่อการเพิ่มประสิทธิภาพของเว็บครอลเลอร์ โดยประเด็นวิจัยหลักสำหรับโครงการนี้ คือการแสดงด้วยการทดลองอย่างเป็นระบบว่า เราสามารถเพิ่มประสิทธิภาพของการครอลเว็บแบบเจาะจงภาษา (language specific crawling) ได้ โดยใช้เว็บคอมมูนิตี้ไมนิ่ง

คำสำคัญ : การครอลเว็บแบบเจาะจงภาษา, เว็บคอมมูนิตี้ไมนิ่ง, ท็อปปิกโมเดล

Research Title: Efficient Crawling of Web 2.0 Applications

Researcher: Dr. Kulwadee Somboonviwat

Faculty: International College **Department:** Software Engineering

ABSTRACT

In recent years, the World Wide Web has become a social platform where the users can easily contribute and publish their information, opinions, and creativity. The term "Web 2.0" was coined in 1999 to describe websites that use technology beyond the static pages of earlier websites. The proliferation of the Web 2.0 has led to the explosive growth of information far beyond a single human mind can consume, leading to the information abundance problem.

This research addresses the Web information abundance problem facing by a web crawler system. The key idea is to apply the web community mining techniques to improving the performance of the web crawler. We empirically show that the feasibilities and benefits of exploiting web community mining in language-specific web crawling.

Keywords : web crawling, web community mining, topic modeling

กิตติกรรมประกาศ

ผู้จัดทำขอขอบพระคุณวิทยาลัยนานาชาติ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ที่ได้ให้การสนับสนุนเงินทุนและสถานที่ในการทำวิจัยนี้ รวมทั้งคณะอาจารย์และเจ้าหน้าที่ของวิทยาลัยฯ ที่ได้ช่วยให้คำแนะนำต่างๆ ที่เป็นประโยชน์ รวมไปถึงอำนวยความสะดวกในด้านต่างๆ ซึ่งมีส่วนช่วยให้การวิจัยนี้สำเร็จตามเป้าหมาย

นอกจากนี้ ผู้วิจัย ขอขอบพระคุณ Assoc.Prof.Dr. Lin Li, Wuhan University of Technology และ Dr. Guandong Xu, University of Technology, Sydney ที่ได้ให้คำแนะนำและความช่วยเหลือที่เป็นประโยชน์อย่างมากในการตีพิมพ์บทความทางวิชาการที่เกี่ยวข้องกับงานวิจัยนี้

การวิจัยครั้งนี้ได้รับทุนสนับสนุนการวิจัยจากสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง จากแหล่งทุน งบประมาณเงินรายได้ ประจำปีงบประมาณ พ.ศ. 2555

ดร. กุลวดี สมบูรณ์วิวัฒน์ หัวหน้าโครงการ

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ก
บทคัดย่อภาษาอังกฤษ ..... ข
กิตติกรรมประกาศ ..... ค
สารบัญ ..... ง
สารบัญตาราง ..... จ
สารบัญภาพ ..... ฉ
บทที่ 1 บทนำ 1
1.1 ความเป็นมาและความสำคัญของปัญหา 1
1.2 วัตถุประสงค์ของการวิจัย 2
1.3 ขอบเขตของการวิจัย 2
1.4 วิธีดำเนินการวิจัย 3
1.5 ประโยชน์ที่คาดว่าจะได้รับ 3
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 4
2.1 เว็บครอลลิ่งแบบเจาะจงภาษา 4
2.2 เว็บคอมมูนิตี้ไมนิ่ง 5
บทที่ 3 วิธีดำเนินการวิจัย 8
3.1 วิธีการวัดความเป็นเนื้อเดียวกันในเชิงภาษาของเว็บคอมมูนิตี้ 8
3.2 ดาต้าเซตที่ใช้ในการทดลอง 9
บทที่ 4 ผลการวิจัย 10
4.1 การกระจายของขนาดของเว็บคอมมูนิตี้ 10
4.2 ความคล้ายคลึงกันในเชิงหัวข้อและความเป็นเนื้อเดียวกันในเชิงภาษาของเว็บคอมมูนิตี้ 11
4.3 การประยุกต์ใช้เว็บคอมมูนิตี้ไมนิ่งในการครอลเว็บแบบเจาะจงภาษา 13
บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ 14
5.1 สรุปผลการวิจัย 14
5.2 ข้อเสนอแนะ 15
บรรณานุกรม **16**
ภาคผนวก 21
ภาคผนวก ก ผลงานวิจัยที่เกี่ยวข้องกับการทำโครงการวิจัยและได้รับการตีพิมพ์เผยแพร่ 21
ประวัตินักวิจัย 22

สารบัญตาราง

ตารางที่ **หน้า**

3.1 คุณสมบัติของ Thai web datasets 9
4.1 ผลการวิเคราะห์ ส่วน C_u ของเว็บคอมมูนิตี้ที่มีค่า $LH_{thai} > 0.5$ 12

สารบัญภาพ

ภาพที่		หน้า
2.1	บล็อคไดอะแกรมของครอลเลอร์แบบเจาะจงภาษา	4
2.2	ตัวอย่างเว็บคอมมูนิตี้ชาร์ต	7
4.1	การกระจายของขนาดของเว็บคอมมูนิตี้	10
4.2	correlation ของ LH_{thai} ที่ได้จาก C_c กับ LH_{thai} ที่ได้จาก C_u	12
4.3	ประเภทของเว็บเพจภายในส่วน uncrawled portion ของเว็บคอมมูนิตี้ที่มีค่า LH_{thai} เกิน 0.5	13

บทที่ 1
บทนำ

บทที่ 1 ของรายงานวิจัย กล่าวถึงความเป็นมาและความสำคัญของปัญหา web information abundance และผลกระทบของปัญหาดังกล่าวต่อเว็บครอลเลอร์ จากนั้นเราจะทำการกำหนดวัตถุประสงค์และขอบเขตของการวิจัย อธิบายภาพรวมของวิธีการดำเนินการวิจัยที่จะใช้ในโครงการวิจัย และสรุปประโยชน์ที่คาดว่าจะได้รับจากการวิจัยนี้

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบัน เวิร์ลไวด์เว็บได้กลายเป็นแพลตฟอร์มของสังคมที่ซึ่งผู้ใช้งานเว็บสามารถสร้างเนื้อหาและแบ่งปันข้อมูล ความคิดเห็นและความคิดสร้างสรรค์ต่างๆ ของตนได้ เว็บ 2.0 เป็นคำศัพท์ที่ได้รับการบัญญัติขึ้นในช่วงปีคริสตศักราช 1999 เพื่อใช้เรียกเทคโนโลยีเว็บยุคใหม่ซึ่งก้าวล้ำกว่าเทคโนโลยีการสร้างเว็บเพจแบบสแตติกในยุคแรก โดยมีคุณลักษณะพิเศษคือ ทำให้ผู้ใช้งานสามารถสร้างและแบ่งปันข้อมูลได้ง่ายดายขึ้น ความแพร่หลายของเทคโนโลยีเว็บ 2.0 รวมไปถึงความสามารถในการเชื่อมต่ออินเตอร์ของผู้ใช้งานที่มากขึ้น ทำให้ข้อมูลเว็บมีปริมาณเพิ่มสูงขึ้นอย่างมหาศาลภายในเวลาอันรวดเร็ว ซึ่งนำไปสู่ปัญหาที่เรียกว่า information abundance ซึ่งหมายถึงการเพิ่มขึ้นอย่างรวดเร็วของปริมาณข้อมูลซึ่งมีขนาดมากเกินกว่าความสามารถของมนุษย์ที่จะซึมซับและนำไปใช้ประโยชน์ได้อย่างมีประสิทธิภาพ

นักวิจัยทั้งภาควิชาการและภาคอุตสาหกรรม ได้พยายามแก้ไขปัญหา information abundance ที่เกิดขึ้นกับผู้ใช้งานเว็บมาตั้งแต่ช่วงแรกของการเกิดขึ้นของเวิร์ลไวด์เว็บ โดยเทคโนโลยีหลักที่ได้ถูกคิดค้นขึ้นมาเพื่อช่วยแก้ไขปัญหาดังกล่าว ได้แก่ เทคโนโลยี เว็บไดเร็กทอรี่ (web directory) และเว็บเสิร์ชเอ็นจิน (web search engine) เป็นต้น

เว็บพอร์ทอลหรือเว็บไดเร็กทอรี่ คือเว็บไซต์ที่รวบรวมลิงค์ไปยังเว็บไซต์ต่างๆ โดยทำการจัดแยกประเภทของลิงค์ออกเป็นหมวดหมู่เพื่อให้ง่ายต่อการสืบค้นและเข้าถึงเว็บไซต์ที่มีเนื้อหาที่ผู้ใช้สนใจ เช่น กีฬา ข่าว บันเทิง เป็นต้น โดยเว็บไดเร็กทอรี่ในยุคแรก เช่น Yahoo!, sanook.com นั้นใช้แรงงานมนุษย์ในการรวบรวมและจัดหมวดหมู่เว็บไซต์ ซึ่งแม้จะทำให้ได้เว็บไดเร็กทอรี่ที่มีเนื้อหาถูกต้อง แต่เนื่องจากปัญหา information abundance ทำให้วิธีการจัดหมวดหมู่ข้อมูลโดยใช้แรงงานมนุษย์ไม่สามารถสร้างหมวดหมู่เว็บไซต์ได้ทันกับปริมาณเว็บไซต์ที่เพิ่มสูงขึ้นอย่างรวดเร็วตลอดเวลาได้ เพื่อแก้ไขปัญหาดังกล่าว นักวิจัยจึงได้เสนอวิธีการทำเหมืองข้อมูลเว็บคอมมูนิตี้หรือ web community mining เพื่อนำมาใช้ในการจัดหมวดหมู่ข้อมูลเว็บปริมาณมหาศาลด้วยโปรแกรมคอมพิวเตอร์ โดยทีมนักวิจัยจากสถาบันวิจัยต่างๆได้เสนออัลกอริธึมสำหรับเว็บคอมมูนิตี้ไม่นิ่ง ขึ้นมาหลากหลายวิธี อาทิเช่น Gibson et al., 1998; Kumar et al., 2999; Flake et al., 2000; Toyoda & Kitsuregawa 2001; Anderson & Lang, 2006 เป็นต้น

เว็บครอลเลอร์ (web crawler) คือซอฟต์แวร์ที่ทำหน้าที่ดาวน์โหลดเว็บเพจปริมาณมากโดยอัตโนมัติจากอินเตอร์เน็ต เว็บครอลเลอร์ได้ถูกนำไปประยุกต์ใช้ในงานด้านต่างๆ มากมาย เช่น การเก็บเว็บเพจสำหรับสร้างดัชนีคำค้นของเว็บเสิร์ชเอ็นจิ้น การสร้างเว็บอาร์ไคว์ (web archive) การมอนิเตอร์เว็บไซต์เพื่อใช้ในการสร้างแอปพลิเคชั่นด้าน business intelligence การเก็บข้อมูลอีเมล์ของผู้ใช้งานอินเตอร์เน็ต ฯ

เนื่องจากเว็บครอลเลอร์ เป็นระบบซอฟต์แวร์ที่ต้องทำหน้าที่ในการเก็บรวบรวมข้อมูลจากเว็บ เพื่อนำข้อมูลที่ได้ไปใช้ในแอปพลิเคชั่นต่างๆ ดังกล่าวข้างต้น เว็บครอลเลอร์จึงต้องสามารถทำงานในสภาพแวดล้อมของเว็บที่มีปริมาณข้อมูลมหาศาลได้อย่างชาญฉลาด กล่าวอีกนัยหนึ่ง เว็บครอลเลอร์จำเป็นต้องชาญฉลาดพอที่จะรับมือกับปัญหา information abundance ในเวิร์ลไวด์เว็บได้ โดยมีเป้าหมายหลักอยู่ที่การใช้ทรัพยากรระบบคอมพิวเตอร์ที่มีอยู่ได้อย่างคุ้มค่าและเหมาะสมกับวัตถุประสงค์ในการนำเว็บครอลเลอร์ไปใช้งาน

เทคโนโลยีเว็บครอลเลอร์ที่ได้รับการพัฒนาขึ้นมาเพื่อแก้ไขปัญหา information abundance คือ การครอลแบบเจาะจง หรือ focused web crawling (Chakrabarti et al., 1999) โดยมีแนวคิดหลักอยู่ที่การใช้เทคนิคทางสถิติ และการเรียนรู้ของเครื่อง (machine learning) เพื่อให้ครอลเลอร์สามารถเลือกครอลเฉพาะเว็บเพจที่เกี่ยวข้องกับหัวข้อที่ผู้ใช้สนใจได้ นอกจากนี้ (Tamura et al., 2007) ยังได้เสนอวิธีการครอลแบบเจาะจงภาษา เพื่อตอบโจทย์ให้กับผู้ใช้งานที่ต้องการรวบรวมข้อมูลเฉพาะในภาษาใดภาษาหนึ่งจากเว็บสเปซ โดยแนวคิดหลักของการครอลแบบเจาะจงภาษาอยู่ที่การใช้ฮิวริสติกที่ได้มาจากการประมวลข้อมูลเว็บกราฟขนาดใหญ่ มาช่วยในการคัดเลือกเว็บเพจที่จะทำการครอล เทคนิคการครอลแบบเจาะจงภาษาดังกล่าวได้ถูกนำไปใช้ในการสร้างเว็บอาร์ไคว์ (web archive) ของเว็บเพจภาษาญี่ปุ่นและภาษาไทย (Tamura et al., 2007; Somboonviwat, 2008)

แม้ว่าการครอลแบบเจาะจงที่เสนอโดย (Tamura et al., 2007) จะมีประสิทธิภาพและสามารถนำไปใช้ในการครอลเว็บในระดับ large-scale ได้จริง แต่วิธีการครอลดังกล่าว ยังมีจุดอ่อนคือไม่รองรับการครอลเว็บเพจที่เขียนด้วยภาษาต่างประเทศที่เกี่ยวข้องกับประเทศที่ใช้ภาษาที่ต้องการ เช่น ข้อมูลเกี่ยวกับประเทศไทยซึ่งถูกเขียนขึ้นโดยใช้ภาษาอังกฤษ เป็นต้น งานวิจัยนี้ต้องการหาแนวทางในการพัฒนาวิธีการเพื่อแก้ไขปัญหา information abundance ที่เกิดขึ้นกับเว็บครอลเลอร์แบบเจาะจงภาษาดังกล่าวมาข้างต้น โดยแนวทางที่โครงการวิจัยนี้ต้องการศึกษา คือการประยุกต์ใช้เทคนิคการทำเว็บคอมมูนิตี้ไมนิ่งเพื่อการเพิ่มประสิทธิภาพของเว็บครอลเลอร์แบบเจาะจงภาษา โดยประเด็นวิจัยหลักสำหรับโครงการนี้ คือการแสดงด้วยการทดลองอย่างเป็นระบบว่า เราสามารถเพิ่มประสิทธิภาพของการครอลเว็บแบบเจาะจงภาษาได้โดยใช้เว็บคอมมูนิตี้ไมนิ่งมาช่วยในการจำแนกลิงค์ที่มีแนวโน้มว่าชี้ไปยังเว็บเพจที่เกี่ยวข้องกับภาษาที่ต้องการครอล

1.2 วัตถุประสงค์ของการวิจัย

โครงการวิจัยนี้ มีวัตถุประสงค์หลักเพื่อศึกษาแนวทางในการนำเทคนิคการทำเว็บคอมมูนิตี้ไมนิ่งไปประยุกต์ใช้ในระบบเว็บครอลเลอร์แบบเจาะจงภาษาเพื่อให้เว็บครอลเลอร์สามารถครอลข้อมูลเว็บเพจที่เกี่ยวข้องกับภาษาที่ผู้ใช้ต้องการให้ได้มากที่สุด และครอลเว็บเพจที่ไม่เกี่ยวข้องกับภาษาที่ผู้ใช้ต้องการให้น้อยที่สุด ทั้งนี้ เพื่อให้วัตถุประสงค์หลักดังกล่าว เราจำเป็นต้องบรรลุวัตถุประสงค์ย่อยต่างๆ ดังต่อไปนี้

๑. สร้างระบบต้นแบบของเว็บครอลเลอร์แบบเจาะจงภาษาเพื่อใช้ในการทดลอง
๒. ศึกษาเทคนิคการทำเว็บคอมมูนิตี้ไมนิ่ง
๓. ทำการทดลองเพื่อศึกษาคุณลักษณะทางด้านภาษาของเว็บเพจที่อยู่ในเว็บคอมมูนิตี้

1.3 ขอบเขตของการวิจัย

เพื่อตอบโจทย์หลักของโครงการวิจัย เราได้กำหนดขอบเขตของการวิจัยไว้ดังต่อไปนี้

๑. การพัฒนาต้นแบบระบบเว็บครอลเลอร์สำหรับเว็บเพจภาษาไทยที่มีประสิทธิภาพ

โครงการวิจัยนี้ จะทำการพัฒนาระบบเว็บครอลเลอร์ต้นแบบ เพื่อนำไปใช้ในการทดลองและการเก็บข้อมูลเว็บเพจภาษาไทย

๒. การทำเว็บคอมมูนิตี้ไมนิ่งที่เหมาะสมกับการนำไปใช้ในระบบเว็บครอลเลอร์สำหรับเว็บเพจภาษาไทย วัตถุประสงค์ในการนำเว็บคอมมูนิตี้ไมนิ่งมาใช้ในระบบเว็บครอลเลอร์แบบเจาะจงภาษา คือเพื่อให้สามารถครอลเว็บเพจที่เกี่ยวข้องได้เพิ่มขึ้น พร้อมกันกับการลดจำนวนการครอลเว็บเพจที่ไม่เกี่ยวข้องให้น้อยลง เพื่อบรรลุวัตถุประสงค์ดังกล่าว เราจำเป็นต้องศึกษาและทดลองหารูปแบบการประยุกต์ใช้เว็บคอมมูนิตี้ไมนิ่งที่เหมาะสมและเข้ากันได้กับระบบการทำงานของเว็บครอลเลอร์แบบเจาะจงภาษาที่ได้ทำการพัฒนาขึ้นในข้อ ๑

1.4 วิธีการดำเนินการวิจัย

การวิจัยในโครงการนี้ จะเน้นการศึกษาด้วยการทำทดลองเป็นหลัก โดยมีขั้นตอนดังต่อไปนี้

๑. สร้างระบบต้นแบบเว็บครอลเลอร์แบบเจาะจงภาษา
๒. สร้างเว็บดาต้าเซตโดยใช้เว็บครอลเลอร์ที่สร้างขึ้นในข้อ ๑
๓. ศึกษาเทคนิคการทำเว็บคอมมูนิตี้ไมนิ่ง
๔. ออกแบบวิธีการวิเคราะห์คุณลักษณะทางด้านภาษาของเว็บคอมมูนิตี้
๕. ทำการทดลองเพื่อวิเคราะห์คุณลักษณะเฉพาะทางด้านภาษาของเว็บคอมมูนิตี้
๖. สรุปผลการทดลองและออกแบบวิธีการหรือรูปแบบที่เหมาะสมในการนำเทคนิคเว็บคอมมูนิตี้ไมนิ่งไปใช้ในระบบเว็บครอลเลอร์แบบเจาะจงภาษา

1.5 ประโยชน์ที่คาดว่าจะได้รับ

๑. ได้ระบบต้นแบบของเว็บครอลเลอร์แบบเจาะจงภาษาเพื่อใช้ในการทดลอง
๒. ได้แนวทางการออกแบบวิธีการทำคอมมูนิตี้ไมนิ่ง สำหรับใช้ในระบบเว็บครอลเลอร์แบบเจาะจงภาษา
๓. ได้เว็บครอลดาต้าเซตเพื่อนำไปใช้ในการทดลองเกี่ยวกับเว็บไมนิ่งต่อไป

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

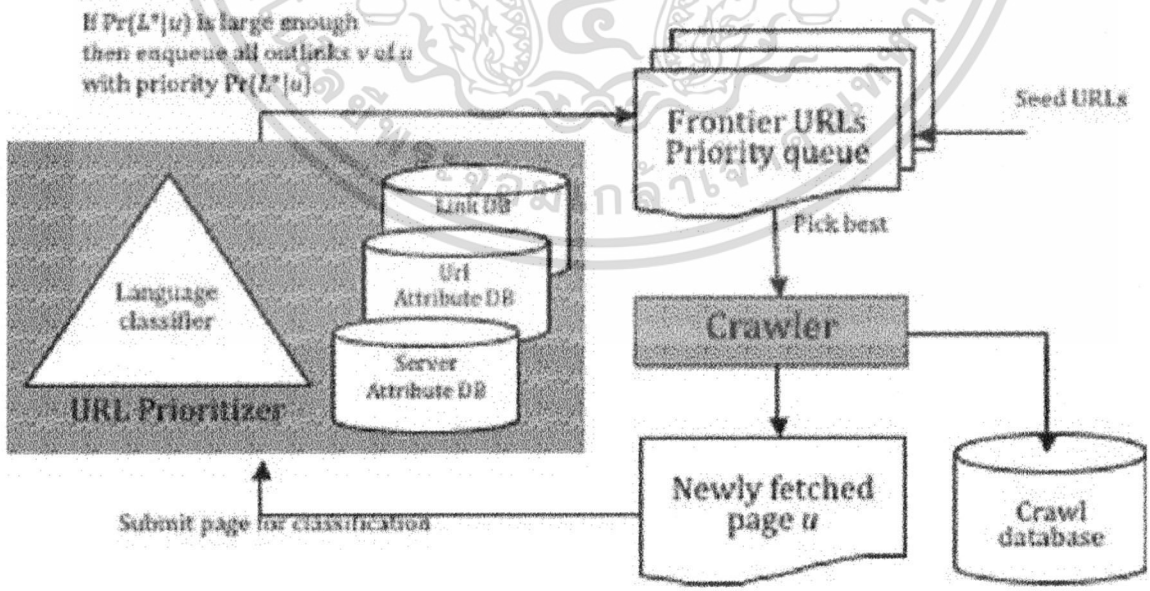
ในบทที่ 2 เราได้รวบรวมทฤษฎีและงานวิจัยต่างๆ และอธิบายในส่วนที่จำเป็นต่อการทำเข้าใจงานวิจัยที่จะนำเสนอในบทต่อๆ ไป โดยทฤษฎีและงานวิจัยที่เกี่ยวข้องกับโครงการวิจัยและจะกล่าวถึงในบทนี้ ได้แก่ ๑) เว็บครอลลิ่งแบบเจาะจงภาษา ๒) เว็บคอมมูนิตี้ไมนิ่ง และ ๓) ท็อปปิกโมเดลลิ่ง

2.1 เว็บครอลลิ่งแบบเจาะจงภาษา

เว็บครอลลิ่งแบบเจาะจงภาษา (language specific crawling; Tamura et al., 2007; Somboonviwat et al. 2006) คือ วิธีการสำหรับการเก็บรวบรวมเว็บเพจที่ถูกสร้างขึ้นด้วยภาษาที่ผู้ใช้งานกำหนด เราสามารถกำหนดนิยามของงานของเว็บครอลเลอร์แบบเจาะจงภาษา (language specific crawler) ได้ดังนี้

คำนิยาม ๑ เว็บครอลเลอร์แบบเจาะจงภาษา. กำหนดเซตของ URL เริ่มต้น และภาษาเป้าหมาย L^* หน้าที่ของเว็บครอลเลอร์แบบเจาะจงภาษาคือการครอลเว็บเพจที่ถูกสร้างขึ้นด้วยภาษา L^* และหลีกเลี่ยงการดาวน์โหลดเว็บเพจที่ถูกสร้างขึ้นด้วยภาษาอื่นให้มากที่สุดเท่าที่จะเป็นไปได้

เฟรมเวิร์กของเว็บครอลเลอร์แบบเจาะจงภาษา มีแนวคิดพื้นฐานมาจากเฟรมเวิร์กของโฟกัสครอลลิ่งซึ่งเสนอโดย Chakrabarti et al., 1999 ภาพที่ 2.1 แสดงบล็อคไดอะแกรมของครอลเลอร์แบบเจาะจงภาษา แนวคิดหลักในการออกแบบระบบเว็บครอลลิ่งแบบเจาะจงภาษา คือ การใช้ตัวจำแนกภาษา (language classifier) เพื่อคาดเภาษาของเว็บเพจที่ครอลเลอร์ได้ดาวน์โหลดมาแล้ว และใช้ความสัมพันธ์ในเชิงภาษาระหว่างเว็บเพจที่มีการเชื่อมต่อกัน เป็นตัวช่วยในการกำหนดทิศทางในการครอลของเว็บครอลเลอร์



ภาพที่ 2.1 บล็อคไดอะแกรมของครอลเลอร์แบบเจาะจงภาษา

2.2 เว็บคอมมูนิตี้ไมนิ่ง

เว็บคอมมูนิตี้คือ เซตของเว็บเพจที่ถูกสร้างขึ้นโดยบุคคลใดบุคคลหนึ่ง หรือกลุ่มคนใดกลุ่มคนหนึ่งซึ่งมีความสนใจเกี่ยวกับหัวข้อเรื่องใดเรื่องหนึ่งร่วมกัน เว็บคอมมูนิตี้จะก่อร่างตัวขึ้นในเว็บ โดยจะอยู่ในรูปของกลุ่มเว็บเพจเกี่ยวกับหัวข้อใดหัวข้อหนึ่งที่มีการเชื่อมโยงกันอย่างหนาแน่น คำนิยามต่างๆ ที่เกี่ยวข้องกับเว็บคอมมูนิตี้ มีดังต่อไปนี้

คำนิยาม ๒ เว็บกราฟ (Web Graph). เว็บกราฟ คือ กราฟแบบมีทิศทาง G(V,E) โดย V คือ เซตของโหนดของเว็บเพจ และ E คือเซตของ ordered pairs (u, v) สำหรับไฮเปอร์ลิงค์จากเว็บเพจ u ไปยังเว็บเพจ v.

คำนิยาม ๓ เว็บคอมมูนิตี้ (Web Community). กำหนดให้เซตแบบจำกัดของเว็บเพจคือ เซต $P=\{p_1, p_2, .., p_n\}$ เว็บคอมมูนิตี้ คือ คู่ $C=(T,M)$ โดย T แทนเซตของหัวข้อเรื่องของคอมมูนิตี้ และ $M \subseteq P$ แทนเซตของเว็บเพจในเซต P ซึ่งเกี่ยวข้องกับหัวข้อ T ถ้า $p_i \in M$ แล้ว เราจะเรียก p_i ว่าเป็นสมาชิกของคอมมูนิตี้ C ขนาดของคอมมูนิตี้เขียนแทนด้วยสัญลักษณ์ $|M|$.

คำนิยาม ๔ การค้นหาเว็บคอมมูนิตี้หรือคอมมูนิตี้ไมนิ่ง (Web Community Discovery and Web Community Mining). กำหนดคอเล็กชั่นของเว็บเพจ P และเว็บกราฟ G ที่สร้างขึ้นจากคอเล็กชั่น P การค้นหาเว็บคอมมูนิตี้หรือคอมมูนิตี้ไมนิ่งคือการสกัดคอมมูนิตี้ k คอมมูนิตี้ $\{C_1,..., C_k\}$ โดย C_i คือ ซับกราฟของ G ซึ่งประกอบด้วยเว็บเพจที่มีเนื้อหาข้อมูลที่มีความสอดคล้องกัน และมีการเชื่อมโยงด้วยไฮเปอร์ลิงค์ระหว่างเว็บเพจภายในที่หนาแน่น

ในช่วงหลายปีที่ผ่านมา ได้มีการเสนอวิธีการค้นหาเว็บคอมมูนิตี้ ขึ้นมามากมายหลายวิธีการ เช่น Gibson et al., 1998; Kleinberg 1998; Kumar 1999; Flake et al., 2000; Toyoda & Kitsuregawa 2006 ในโครงการวิจัยนี้เราจะทำการทดลองโดยใช้วิธีการค้นหาเว็บคอมมูนิตี้ที่เสนอโดย Toyoda & Kitsuregawa 2001 รายละเอียดของวิธีการค้นหาเว็บคอมมูนิตี้ของ Toyoda & Kitsuregawa, 2001 สามารถอธิบายได้ดังต่อไปนี้

2.2.1 อัลกอริธึมสำหรับค้นหาเว็บคอมมูนิตี้ที่เสนอโดย Toyoda & Kitsuregawa, 2001

Toyoda & Kitsurewaga, 2001 ได้เสนอเทคนิคสำหรับสร้างเว็บคอมมูนิตี้ชาร์ต (web community chart) ซึ่งสามารถค้นหาได้ทั้งเว็บคอมมูนิตี้และความสัมพันธ์ระหว่าเว็บคอมมูนิตี้ที่ได้สกัดออกมา เทคนิคดังกล่าวได้แนวความคิดมาจาก related page algorithm ซึ่งเป็นอัลกอริธึมที่ใช้สำหรับหาความเกี่ยวข้องกันของเว็บเพจในเซตที่กำหนด โดยมีหลักการคือการตรวจสอบ derivation relationship ระหว่างเว็บเพจ ซึ่งสามารถอธิบายได้ดังต่อไปนี้ ถ้าเว็บเพจ s derives เว็บเพจ t โดยความสัมพันธ์ related page และ เว็บเพจ t derives เว็บเพจ s โดยความสัมพันธ์ related page แล้ว เราจะกล่าวว่าเว็บเพจ s และเว็บเพจ t มีความสัมพันธ์กันแบบ symmetric derivation relationship ตัวอย่างเช่น แฟนเพจ i ของทีมเบสบอล derives แฟนเพจ j <> i โดยความสัมพันธ์ related page และเพจ j ก็ derives แฟนเพจ i โดยความสัมพันธ์ related page ดังนั้นเรากล่าวว่า แฟนเพจ i และ j มีความสัมพันธ์กันแบบ symmetric derivation relationship อนึ่ง โดยส่วนมาก เว็บเพจที่มีความสัมพันธ์แบบ symmetric derivation relationship จะเป็นเว็บเพจที่ถูกชี้โดยกลุ่มของฮับเพจที่คล้ายกัน

กรณีที่เพจ s derives เพจ t โดยความสัมพันธ์ related page แต่ว่าเพจ t ไม่ได้ derives เพจ s โดยความสัมพันธ์ related page ซึ่งหมายความว่า เพจ t เป็นเว็บเพจที่ถูกชี้โดยฮับเพจหลายๆ เพจ ยกตัวอย่างเช่น แฟนเพจของทีมเบสบอล มีความสัมพันธ์ related page กับเพจอย่างเป็นทางการของทีม แต่เพจอย่างเป็นทางการของทีมจะไม่มีความสัมพันธ์ related page กับเพจแฟนเพจเหล่านั้น ในกรณีนี้เรา

จะกล่าวได้ว่า เพจอย่างเป็นทางการของทีมเบสบอลมีความสัมพันธ์กับเว็บคอมมูนิตี้ของแฟนเพจ แต่เพจอย่างเป็นทางการของทีมเบสบอลดังกล่าวเป็นสมาชิกของเว็บคอมมูนิตี้ของทีมเบสบอล ความสัมพันธ์ derivation relationship ดังกล่าวมานี้ จะถูกนำมาใช้ในการออกแบบอัลกอริธึมสำหรับค้นหาเว็บคอมมูนิตี้ที่เกี่ยวข้องกัน (related communities) ต่อไป

โดยสรุป การระบุเว็บคอมมูนิตี้และความสัมพันธ์ระหว่างเว็บคอมมูนิตี้ สามารถกระทำได้ด้วยการใช้ความสัมพันธ์ derivation relationship ระหว่างเว็บเพจใดๆ กับ เว็บเพจที่มีความเกี่ยวข้องกัน โดยเรานิยามเว็บคอมมูนิตี้ ว่าเป็นกลุ่มของเพจที่เชื่อมโยงกันด้วยความสัมพันธ์แบบ symmetric derivation relationship อย่างหนาแน่น และคอมมูนิตี้สองคอมมูนิตี้ใดๆ จะสัมพันธ์กันก็ต่อเมื่อสมาชิกของคอมมูนิตี้หนึ่ง derives เว็บเพจสมาชิกของคอมมูนิตี้อีกคอมมูนิตี้หนึ่ง

จากแนวความคิดดังกล่าว Toyoda & Kitsuregawa, 2001 ได้เสนออัลกอริธึมสำหรับสร้างเว็บคอมมูนิตี้ชาร์ต (web community chart construction algorithm) ซึ่งสามารถอธิบายได้ดังต่อไปนี้

กำหนด เซตของเว็บเพจเริ่มต้น (seed set) ขั้นตอนการสร้างเว็บคอมมูนิตี้ชาร์ต สามารถอธิบายได้ดังต่อไปนี้

(1) ขยายเซตเริ่มต้น

เนื่องจากขนาดของเซตของเว็บเพจเริ่มต้น หรือ seed set มีขนาดเล็กเกินไปสำหรับการสกัด symmetric relationship ดังนั้นเราจึงจำเป็นต้องขยายเซตเริ่มต้นก่อน โดยใช้อัลกอริธึม related page กับเว็บเพจแต่ละเพจที่อยู่ในเซตเริ่มต้น และเลือก top N authorities ที่ได้จากผลลัพธ์แต่ละผลลัพธ์ เพิ่มเข้าไปเป็นสมาชิกของเซตเริ่มต้น สำหรับอัลกอริธึม related page สามารถอ้างอิงได้จาก Toyoda & Kitsuregawa, 2001.

(2) สร้าง authority derivation graph

Authority derivation graph (ADG) คือ กราฟแบบมีทิศทางซึ่งแสดงความสัมพันธ์ derviation relationships ระหว่างเพจในเซตเริ่มต้น แต่ละ node ใน ADG คือ เพจในเซตเริ่มต้นที่ได้ถูกขยายแล้วโดยวิธีการในข้อ (1) และแต่ละ directed edge ใน ADG จากโหนด s ไปยังโหนด t ใช้แทนความสัมพันธ์ s derives t โดย t อยู่ใน top N authorizes (N=10) ของผลลัพธ์ที่ได้จาก related page algorithm: Companion-.

(3) สกัด symmetric derivation graph

Symmetric derivation graph (SDG) คือกราฟแบบมีทิศทาง ซึ่งแสดงความสัมพันธ์ mutual derivation relationship ระหว่างเพจในเซตเริ่มต้นที่ถูกขยายแล้ว โหนดใน SDG คือโหนดใน ADG และถ้าเพจ s และเพจ t ใน ADG มีการเชื่อมต่อซึ่งกันและกันแล้ว จะมี directed edge จากโหนด s ไปยังโหนด t ใน SDG

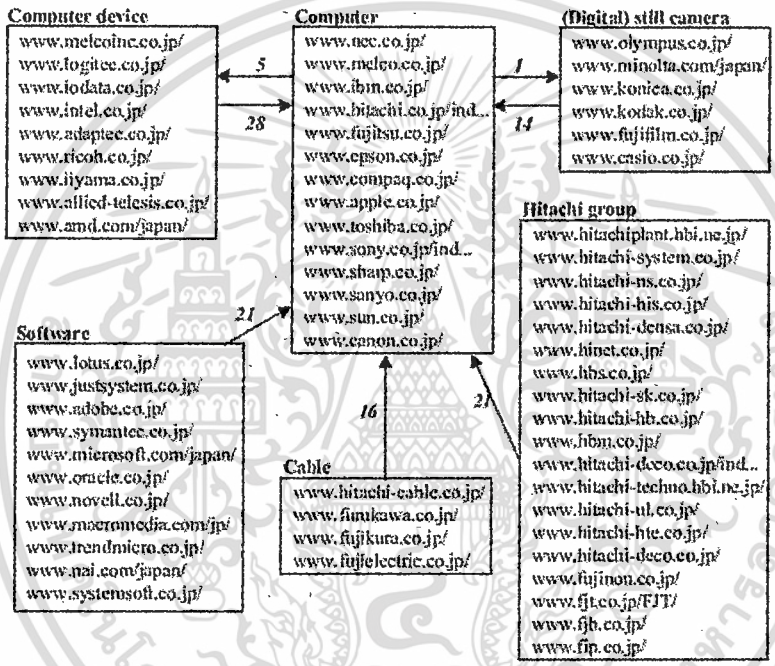
(4) ระบุ web communities

ในขั้นตอนนี้เราจะทำการระบุเว็บคอมมูนิตี้จากกราฟ SDG ที่ได้ในข้อ (3) โดยการแบ่งกราฟด้วย node triangle ตามขั้นตอนดังต่อไปนี้

1. ระบุตำแหน่งของสามเหลี่ยมของโหนดใน SDG ใช้สามเหลี่ยมเป็น community core
2. เพิ่มโหนดทุกโหนดที่ถูกลิงก์โดยโหนดที่อยู่ใน community core
3. จากข้อ 2 เราจะได้ web community partition หนึ่ง partition ให้ลบ partition ดังกล่าวออกจาก SDG
4. ทำข้อ 1-3 ใหม่จนกว่าเซตของโหนดของ SDG จะเป็นเซตว่าง

(5) สร้าง web community chart

เว็บคอมมูนิตี้ชาร์ต คือ กราฟแบบมีทิศทาง ซึ่งโหนดในกราฟดังกล่าว คือ คอมมูนิตี้ และ weighted directed edge ซึ่งเชื่อมต่อคอมมูนิตี้ที่เกี่ยวข้องกัน เว็บคอมมูนิตี้ชาร์ตสามารถสร้างได้โดยเริ่มจากเซตของคอมมูนิตี้ที่ระบุได้ในข้อ (4) จากนั้น edges ระหว่างคอมมูนิตี้สามารถระบุได้โดยใช้ข้อมูลจาก ADG ดังนี้คือ เราจะเพิ่ม directed edge ที่มีค่าถ่วงน้ำหนักเป็น w จากคอมมูนิตี้ c ไปยังคอมมูนิตี้ d <> c ก็ต่อเมื่อ มี directed edge w ใน ADG จากโหนดในคอมมูนิตี้ c ไปยังโหนดในคอมมูนิตี้ d ภาพที่ 2.2 แสดงตัวอย่างของส่วนหนึ่งของเว็บคอมมูนิตี้ชาร์ตที่ได้จากอัลกอริธึมนี้



ภาพที่ 2.2 ตัวอย่างเว็บคอมมูนิตี้ชาร์ต

บทที่ 3
วิธีดำเนินการวิจัย

ในบทที่ 3 นี้เราจะอธิบายวิธีการดำเนินการวิจัย ที่ผู้วิจัยได้คิดค้นขึ้นเพื่อนำไปใช้ในการวิเคราะห์คุณลักษณะในเชิงความเป็นเนื้อเดียวกันของภาษาของเว็บเพจในเว็บคอมมูนิตี้ นอกจากนี้ เรายังอธิบายดาต้าเซตที่จะนำมาใช้ในการทดลองด้วย

3.1 วิธีการวัดความเป็นเนื้อเดียวกันในเชิงภาษาของเว็บคอมมูนิตี้

ในโครงการนี้ ผู้วิจัยได้เสนอตัววัดค่าสองชนิดเพื่อนำมาใช้ในการวัดคุณลักษณะความเป็นเนื้อเดียวกันในเชิงภาษาของเว็บคอมมูนิตี้ สองค่าด้วยกัน คือ topical similarity with topics in web directory และ language homogeneity

คำนิยาม ๕ ความคล้ายคลึงกันด้านเนื้อความ (Topical similarity). กำหนดให้ C=(T,M) เป็นเว็บคอมมูนิตี้เกี่ยวกับหัวข้อ T ซึ่งสมาชิกของ C ประกอบด้วยเว็บเพจในเซต M = {m_1, m_2, ..., m_n} และกำหนดให้ TD = {s_1, s_2, ..., s_p} เป็น topic directory ที่มีประกอบไปด้วยหมวดหมู่ต่างๆ จำนวน p หมวดหมู่ และแต่ละหมวดหมู่ s_j, j=1...p จะมีสมาชิกเป็นเซตของเว็บเพจ U_{sj} = {$u_{sj,1}$, $u_{sj,2}$, ..., $u_{sj,q}$} เราสามารถคำนวณค่า topical similarity (TS) ระหว่างคอมมูนิตี้ C กับหมวดหมู่ s_j ได้ดังสมการที่ 1

$$TS(C,s_j)=\frac{|M \cap U_{s_j}|}{|M|} \qquad (1)$$

ในที่นี้ |M| คือจำนวนของเว็บเพจสมาชิกในคอมมูนิตี้ $|M \cap U_{sj}|$ คือจำนวนของสมาชิกในเซตที่ได้จากการอินเตอร์เซ็กชั่นกันระหว่างเซต M และเซต U_{sj} และ $0 \le TS(C,sj) \le 1$

คำนิยาม ๖ ความเป็นเนื้อเดียวกันในเชิงภาษา (Language Homogeneity). กำหนดให้ C=(T,M) เป็นเว็บคอมมูนิตี้เกี่ยวกับหัวข้อ T ซึ่งสมาชิกของ C ประกอบด้วยเว็บเพจในเซต M = {m_1, m_2, ..., m_n} และกำหนดให้ l เป็นชื่อของภาษาที่ต้องการ และ $M_l \subseteq M$ เป็นเซตของเว็บเพจใน M ซึ่งถูกสร้างขึ้นโดยใช้ภาษา l แล้ว ค่าของ language homogeneity (LH) ของคอมมูนิตี้ C เทียบกับภาษา l สามารถคำนวณได้ดังสมการที่ 2

$$LH_l=\frac{|M_l|}{|M|} \qquad (2)$$

ในที่นี้ $|M_l|$ คือจำนวนของเว็บเพจสมาชิกในคอมมูนิตี้ซึ่งถูกสร้างขึ้นโดยใช้ภาษา l และ $0 \le LH_l \le 1$

3.2 ดาต้าเซตที่ใช้ในการทดลอง

ดาต้าเซตที่ใช้ในโครงการนี้ คือ Thai web datasets ซึ่งสร้างใน Somboonviwat, 2008. โดยใช้ BFS crawler และ language specific crawler ตารางที่ 3.1 สรุปคุณสมบัติด้านต่างๆ ของ Thai web datasets สำหรับรายละเอียดต่างๆ เกี่ยวกับขั้นตอนการสร้างเว็บดาต้าเซตนี้ สามารถอ้างอิงได้จาก Somboonviwat, 2008

ตารางที่ 3.1 คุณสมบัติของ Thai web datasets

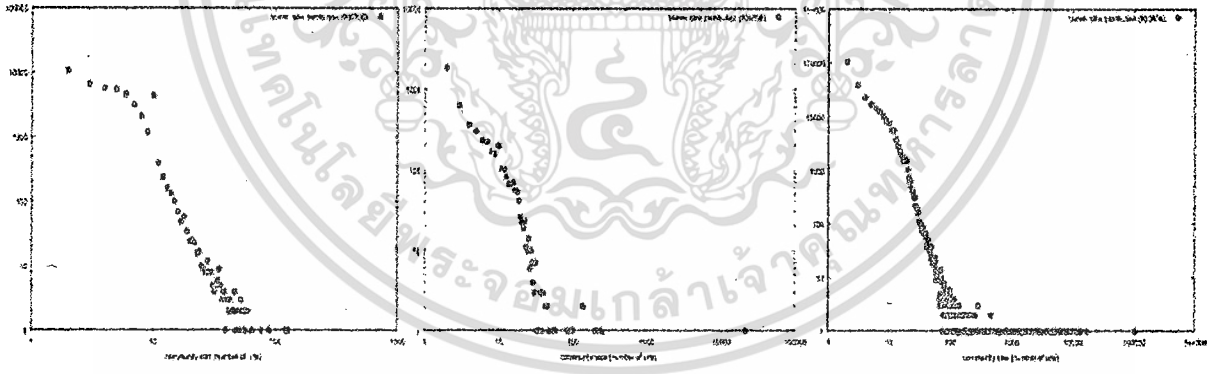
Dataset	200705	200706	200707	200708	200804
Crawling method	Language Specific	Language Specific	Language Specific	Language Specific	BFS
Number of HTML pages	1,188,217	2,063,858	2,451,383	2,295,208	68,839,605
Thai Language HTML pages	502,909	420,074	635,450	602,528	3,633,975
Ratio of Thai pages	0.42	0.20	0.25	0.26	0.05
Number of communities	45,443	6,199	6,781	4,977	265,476

บทที่ 4
ผลการวิจัย

ในบทที่ 4 นี้ เราจะรายงานผลการวิจัยเกี่ยวกับการวิเคราะห์คุณลักษณะทางด้านภาษาของเว็บคอมมูนิตี้ โดยผลการวิจัยที่น่าสนใจ ที่เราจะรายงานในที่นี้ ได้แก่ การกระจายของขนาดของเว็บคอมมูนิตี้ในหัวข้อ 4.1 ความคล้ายคลึงกันในเชิงหัวข้อเรื่องและความเป็นเนื้อเดียวกันในเชิงภาษาของเว็บคอมมูนิตี้ในหัวข้อ 4.2 จากนั้นเราจะนำผลการวิเคราะห์ที่ได้จากข้อ 4.1 และ 4.2 เพื่อนำเสนอการประยุกต์ใช้เทคนิคการทำเว็บคอมมูนิตี้ไมนิ่งในการครอลเว็บแบบเจาะจงภาษาในหัวข้อ 4.3 ต่อไป

4.1 การกระจายของขนาดของเว็บคอมมูนิตี้

เราได้ทำการสกัดเว็บคอมมูนิตี้จาก Thai web datasets (อ้างอิงหัวข้อ 3.2) โดยใช้อัลกอริธึมสำหรับสร้าง web community chart ในหัวข้อ 2.2.1 จากเว็บคอมมูนิตี้ชาร์ตที่ได้ เราทำการวัดขนาดของเว็บคอมมูนิตี้แต่ละคอมมูนิตี้ แล้วสร้างกราฟแบบ log-log plot ที่แสดงการกระจายของขนาดของเว็บคอมมูนิตี้ที่สกัดออกมาได้จากเว็บดาต้าเซต 200705, 200708 และ 200804 โดยเราแสดงผลที่ได้ในรูปภาพที่ 4.1 จากรูปกราฟ เราสามารถสังเกตได้ว่า การกระจายของขนาดของคอมมูนิตี้ของดาต้าเซตทุกดาต้าเซตสอดคล้องกับกฎการกระจายแบบ power-law distribution ซึ่งเป็นคุณลักษณะที่สำคัญอย่างหนึ่งที่พบได้ในเว็บกราฟ โดยคอมมูนิตี้ที่พบส่วนใหญ่จะเป็นคอมมูนิตี้ขนาดเล็กซึ่งประกอบด้วยสมาชิกน้อยกว่า 10 เว็บเพจ อย่างไรก็ตามเราพบว่า มีเว็บคอมมูนิตี้ขนาดใหญ่อยู่ประมาณสองถึงสามคอมมูนิตี้ในแต่ละดาต้าเซต



(a) 200705 (b) 200708 (c) 200804

ภาพที่ 4.1 การกระจายของขนาดของเว็บคอมมูนิตี้

นอกจากนี้ เรายังได้ทำการตรวจสอบเกี่ยวกับเว็บเพจในเว็บคอมมูนิตี้ที่มีขนาดใหญ่มาก และพบว่า ส่วนใหญ่จะเป็นเว็บเพจที่เกี่ยวข้องกับ blog websites, spam pages, หรือ pornographic websites เนื่องจากเว็บเพจเหล่านี้ ไม่เกี่ยวข้องกับเนื้อความที่เราสนใจ ในโครงการนี้ เราจะตัดเว็บคอมมูนิตี้เหล่านี้ออกไปจากการวิเคราะห์ในส่วนต่อไป

4.2 ความคล้ายคลึงกันในเชิงหัวข้อและความเป็นเนื้อเดียวกันในเชิงภาษาของเว็บคอมมูนิตี้

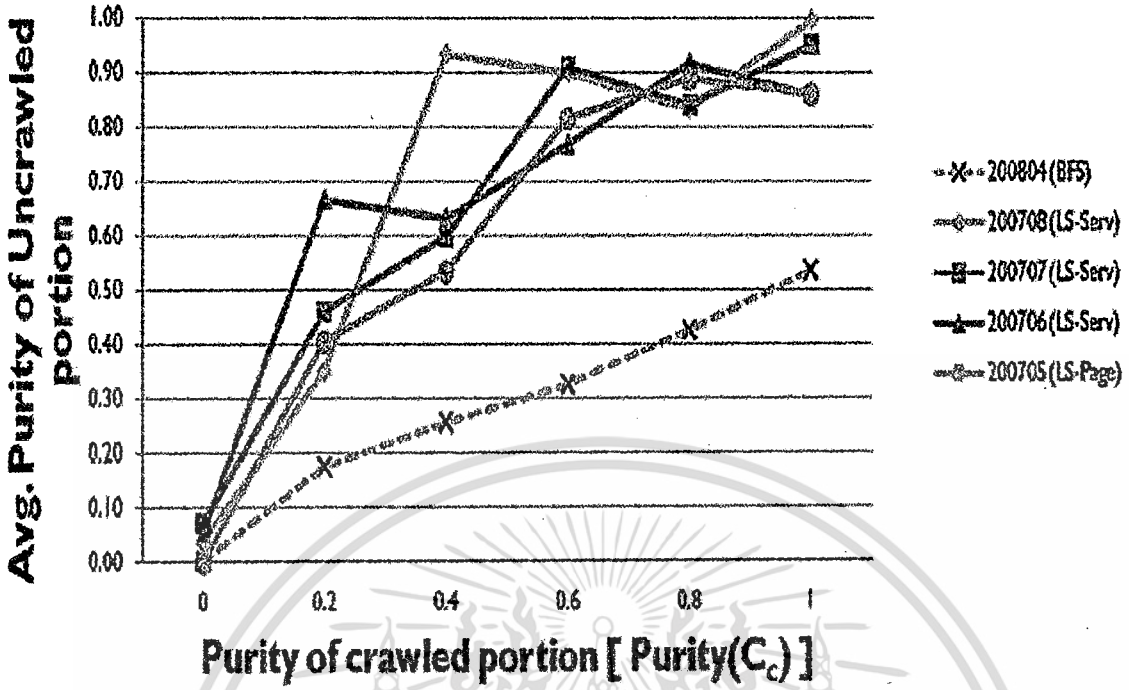
การทดลองในหัวข้อนี้ มีวัตถุประสงค์เพื่อวัดค่าของ topical similarity โดยเราจะเทียบความคล้ายคลึงกันในเชิงหัวข้อระหว่างเว็บคอมมูนิตี้ที่สกัดได้จาก Thai web datasets กับเว็บไดเร็กทอรี่ของ http://www.truehits.net/ โดยเราได้ทำการดาวน์โหลดเว็บไดเร็กทอรี่ดังกล่าว ซึ่งประกอบไปด้วย ลิงก์จำนวน 10,000 ลิงก์ ที่ถูกจัดหมวดหมู่อยู่ใน 19 หมวดหมู่ เราทำการ sample เว็บคอมมูนิตี้มาจำนวนหนึ่งและคำนวณหาค่า TS โดยใช้สมการที่ 1 ในบทที่ 3 ผลลัพธ์ที่ได้ชี้ให้เห็นว่า เว็บคอมมูนิตี้ที่สกัดได้กับเว็บไดทอรี่ที่ถูกสร้างขึ้นด้วยวิธีแบบ manual นั้นมีความคล้ายคลึงกันในเชิงหัวข้อสูงมาก คืออยู่ที่ระดับค่า TS เท่ากับ 0.9

ในส่วนของความเป็นเนื้อเดียวกันในเชิงภาษา (language homogeneity) เราพบว่าคอมมูนิตี้ที่สกัดได้จากดาต้าเซตที่สร้างขึ้นด้วยวิธี language specific crawling จะมีความเป็นเนื้อเดียวกันในเชิงภาษาที่สูงกว่าเว็บคอมมูนิตี้ที่สร้างขึ้นด้วยวิธี breadth-first crawling

เพื่อทำความเข้าใจเกี่ยวกับคุณลักษณะของข้อมูลในแต่ละดาต้าเซตให้มากขึ้น เราได้ทำการทดลองเพื่อวัดค่าของ LH_{thai} โดยในการทดลองนี้เราแบ่งเซตของ URL ภายในคอมมูนิตี้แต่ละคอมมูนิตี้ออกเป็นสองส่วน (two disjoint sets) ส่วนแรก เป็นเซตของ URL ของเว็บเพจที่ครอลเลอร์ได้ดาวน์โหลดมาเรียบร้อยแล้ว กำหนดชื่อเป็น C_c ส่วนที่สอง เป็นเซตของ URL ของเว็บเพจที่ครอลเลอร์ยังไม่ได้ดาวน์โหลด กำหนดชื่อเป็น C_u จากนั้น ทำการคำนวณค่า LH_{thai} ของ C_c และ C_u ของแต่ละคอมมูนิตี้ และพลอตกราฟ เพื่อหา correlation ระหว่างค่าของ LH_{thai} ของ C_c และ C_u ผลลัพธ์ของการทดลองนี้แสดงอยู่ในภาพที่ 4.2 และตารางที่ 4.1

จากภาพที่ 4.2 และตารางที่ 4.1 เราสามารถสรุปเป็นข้อสังเกตได้ดังต่อไปนี้

- เว็บคอมมูนิตี้ที่ถูกสกัดออกมาจากดาต้าเซตที่สร้างโดยใช้ language specific crawler จะมีความเป็นเนื้อเดียวกันในเชิงภาษาที่สูง ซึ่งหมายความว่า language specific crawler มีความสามารถในการรวบรวมเว็บเพจแบบเจาะจงภาษาที่สูงกว่าเว็บครอลเลอร์แบบทั่วไปหรือ BFS crawler
- มีความสัมพันธ์ correlation ที่สูงระหว่าง ค่าของ LH_{thai} ของ C_c และ C_u ในแต่ละคอมมูนิตี้ โดยเฉพาะอย่างยิ่ง ในกรณีที่ดาต้าเซตถูกสร้างขึ้นด้วย language specific crawler โดยเราสามารถสรุปเป็นข้อสังเกตได้ ดังนี้คือ ถ้าค่าของ LH_{thai} ของ C_c มีค่าสูงแล้วจะมีความเป็นไปได้มากที่ค่าของ LH_{thai} ของ C_u ก็จะมีค่าสูงด้วย
- จากข้อสังเกตข้างต้น เราตั้งสมมติฐานว่า ส่วน C_u ของเว็บคอมมูนิตี้ที่มีค่า LH_{thai} มากกว่า 0.5 สามารถนำไปใช้ช่วยเพิ่มประสิทธิภาพของ language specific crawler ได้



ภาพที่ 4.2 correlation ของ LH_{thai} ที่ได้จาก C_c กับ LH_{thai} ที่ได้จาก C_u

ตารางที่ 4.1 ผลการวิเคราะห์ ส่วน C_u ของเว็บคอมมูนิตี้ที่มีค่า $LH_{thai} > 0.5$

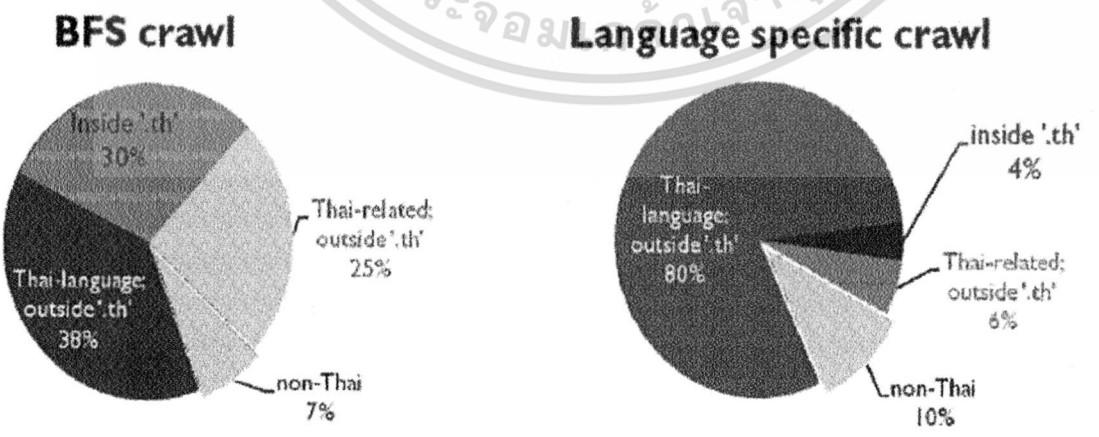
Dataset	200705	200706	200707	200708	200804
Crawling method	Language Specific	Language Specific	Language Specific	Language Specific	BFS
Number of uncrawled pages	155	233	201	170	439
Number of uncrawled thai pages	126	194	183	164	165
Ratio of thai pages in the uncrawled portion	0.81	0.83	0.91	0.96	0.38

4.3 การประยุกต์ใช้เว็บคอมมูนิตี้ไมนิ่งในการครอลเว็บแบบเจาะจงภาษา

ในหัวข้อนี้เราจะอธิบายแนวความคิดในการประยุกต์ใช้เทคนิค language homogeneity analysis เพื่อเพิ่มประสิทธิภาพของ language specific crawler แนวคิดหลักของเราอยู่ที่การคาดเดาภาษาของเว็บเพจที่ครอลเลอร์ยังไม่ได้ดาวน์โหลดในเว็บคอมมูนิตี้หนึ่ง โดยใช้ค่า language homogeneity ของเว็บเพจที่ครอลเลอร์ดาวน์โหลดมาแล้วที่อยู่ในเว็บคอมมูนิตี้นั้น

จากผลการวิเคราะห์เว็บคอมมูนิตี้ที่ได้กล่าวไปในหัวข้อก่อนหน้า เราพบว่า ในเว็บคอมมูนิตี้ใดๆ ถ้าเว็บเพจที่ครอลเลอร์ดาวน์โหลดมาแล้วส่วนใหญ่เป็นภาษาไทย แล้วจะมีแนวโน้มสูงที่ URL ในส่วนของเว็บคอมมูนิตี้ที่ครอลเลอร์ยังไม่ได้ดาวน์โหลดจะชี้ไปยังเว็บเพจที่เกียวข้องกับภาษาไทย เราสามารถนำผลการวิเคราะห์ดังกล่าวไปใช้เพิ่ม URL ในเซตเริ่มต้น (seed URL set) ของเว็บครอลเลอร์ได้ โดยการเพิ่ม uncrawled URLs ที่ได้จากเว็บคอมมูนิตี้ที่มีค่า language homogeneity สูงกว่า 0.5 เข้าไปในเซตเริ่มต้นของ language specific crawler

เพื่อแสดงให้เห็นความเป็นไปได้ของแนวความคิดข้างต้น เราได้ทำการศึกษาประเภทของเว็บเพจที่อยู่ใน uncrawled portions ของเว็บคอมมูนิตี้ที่มีค่า language homogeneity (LH_{thai}) มากกว่า 0.5 ผลการทดลองที่ได้แสดงในภาพที่ 4.3 โดยภาพดังกล่าวแสดงการจำแนกประเภทของเว็บเพจซึ่งอยู่ใน uncrawled portions ซึ่งจะเห็นได้ว่ากว่า 90% ของ URL ใน uncrawled portion เป็นเว็บเพจที่มีความเกี่ยวข้องกับภาษาไทย นอกจากนี้ เรายังพบว่า มีเว็บเพจจำนวนมากในส่วน uncrawled portion ที่ถูกสร้างขึ้นด้วยภาษาต่างประเทศที่ไม่ใช่ภาษาไทยแต่เป็นเว็บเพจที่มีเนื้อความเกี่ยวข้องกับประเทศไท ผลลัพธ์ดังกล่าว ชี้ให้เห็นว่า การนำแนวคิดดังกล่าวไปผนวกเข้ากับเทคโนโลยีของ language specific crawler ที่มีอยู่ จะทำให้ครอลเลอร์สามารถทำหน้าที่เก็บรวบรวมเว็บเพจได้ทั้งที่เขียนขึ้นด้วยภาษาไทย และเขียนขึ้นด้วยภาษาต่างประเทศแต่มีความเกี่ยวข้องในด้านเนื้อหากับประเทศไทย ซึ่งหากเราสามารถสร้างครอลเลอร์ที่มีความสามารถเช่นนี้ได้ เราก็จะสามารถนำระบบครอลเลอร์ดังกล่าวไปใช้ในการสร้างเว็บอาร์ไคว์แห่งชาติได้ (national web archive) โดยเว็บอาร์ไคว์ดังกล่าวจะสามารถนำไปใช้ประโยชน์ได้หลายทาง อาทิเช่น ใช้เพื่อเก็บรวบรวมประวัติศาสตร์เว็บไทย ใช้เพื่อการศึกษาในเชิงสังคมศาสตร์ ภาษาศาสตร์ และวัฒนธรรม เป็นต้น



ภาพที่ 4.3 ประเภทของเว็บเพจภายในส่วน uncrawled portion ของเว็บคอมมูนิตี้ที่มีค่า LH_{thai} เกิน 0.5

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

ในบทที่ 5 นี้ เราจะสรุปผลที่ได้จากการวิจัยในโครงการนี้ รวมทั้งให้ข้อเสนอแนะเกี่ยวกับการนำผลลัพธ์ที่ได้จากงานวิจัยไปใช้ประโยชน์ รวมถึงเสนอแนวทางในการวิจัยต่อยอดในอนาคต

5.1 สรุปผลการวิจัย

โครงการวิจัยนี้ ได้ทำการศึกษาเกี่ยวกับการครอลเว็บแบบเจาะจงภาษา โดยโจทย์วิจัยหลักของโครงการนี้ คือการเพิ่มประสิทธิภาพในการรวบรวมเว็บเพจที่เกี่ยวข้องกับภาษาที่ผู้ใช้งานต้องการ โดยประสิทธิภาพในที่นี้ หมายถึงการที่เว็บครอลเลอร์ สามารถดาวน์โหลดเว็บเพจที่เกี่ยวข้องกับภาษาที่ผู้ใช้ต้องการได้เป็นจำนวนมากขึ้น และขณะเดียวกันกับที่การดาวน์โหลดเว็บเพจที่ไม่เกี่ยวข้องกับภาษาที่ผู้ใช้ต้องการมีจำนวนน้อยลง

ซึ่งในโครงการวิจัยนี้ เราสนใจศึกษาการประยุกต์ใช้เทคโนโลยีเว็บคอมมูนิตี้ไมนิ่งเพื่อการเพิ่มประสิทธิภาพของเว็บครอลเลอร์แบบเจาะจงภาษาหรือ language specific crawler โดยเพื่อให้บรรลุวัตถุประสงค์ดังกล่าว เราได้ทำการศึกษาวิจัยเพื่อตอบปัญหา ดังต่อไปนี้ คือ

(1) ความเป็นเนื้อเดียวกันของภาษาของเว็บเพจในคอมมูนิตี้
เราพบว่าเว็บเพจในคอมมูนิตี้ที่ได้จากดาต้าเซตที่ถูกสร้างขึ้นโดยใช้ language specific crawler จะมีความเป็นเนื้อเดียวกันของภาษาของเว็บเพจที่สูง

(2) ความสัมพันธ์ระหว่างคุณลักษณะในเชิงภาษาของเว็บเพจในคอมมูนิตี้ที่ครอลเลอร์ดาวน์โหลดมาแล้วกับที่ครอลเลอร์ยังไม่ได้ดาวน์โหลด
เราพบว่า ถ้าเว็บเพจส่วนที่ครอลเลอร์ดาวน์โหลดมาแล้วในคอมมูนิตี้หนึ่ง ส่วนใหญ่เป็นเว็บเพจที่มีเนื้อหาเป็นภาษาไทย แล้วเว็บเพจส่วนที่ครอลเลอร์ยังไม่ได้ดาวน์โหลด ก็จะมีแนวโน้มสูงที่จะเป็นเว็บเพจที่เขียนด้วยภาษาไทย หรือมีเนื้อหาเกี่ยวข้องกับเว็บไทย

(3) ความเป็นไปได้ในการนำความสัมพันธ์ในข้อสองข้างต้น ไปใช้ในการเพิ่มประสิทธิภาพของครอลเลอร์แบบเจาะจงภาษา (language specific crawler)
เราได้ทำการศึกษาและแสดงผลการทดลองที่ชี้ให้เห็นว่า การเพิ่มประสิทธิภาพการรวบรวมเว็บเพจของ language specific crawler สามารถทำได้ โดยการค้นหาเว็บคอมมูนิตี้ที่มีความเป็นเนื้อเดียวกันในเชิงภาษาที่สูง และนำลิงก์ที่อยู่ในเว็บคอมมูนิตี้ดังกล่าว ที่ครอลเลอร์ยังไม่ได้ดาวน์โหลด หรือยังไม่เคยพบ ไปเพิ่มเข้าในเซตเริ่มต้นของครอลเลอร์

5.2 ข้อเสนอแนะ

นอกจากผลการวิจัยที่ได้ ดังสรุปไว้ในหัวข้อ 5.1 ผู้วิจัยพบว่า การวิจัยในโครงการนี้ ยังทำให้เกิดความรู้เพิ่มเติม ดังต่อไปนี้

(1) การครอลเว็บ

หลักการทำงานของเว็บครอลเลอร์ แม้ว่าจะไม่มีความซับซ้อนมาก แต่ในการปฏิบัติจริงกับพบปัญหาต่างๆ ทั้งในเชิงเทคนิค และในเชิงสังคมมากมาย เช่น การครอลโดยใช้เครือข่ายที่เป็นของส่วนรวม เราจำเป็นต้องคำนึงถึงผู้ใช้คนอื่น และในขณะเดียวกันเราก็จำเป็นต้องใช้ทรัพยากรเครือข่ายให้คุ้มค่าที่สุด

(2) การวิเคราะห์ภาษาของเว็บเพจ

การทำนายภาษาของเว็บเพจด้วย language classifier เราใช้ API ของเว็บบราวเซอร์ Mozilla และจำเป็นต้องมีการทดลองเพื่อ train language classifier รวมทั้งต้องมีการ preprocess เว็บเพจก่อนที่จะป้อนให้กับ language classifier มิฉะนั้นผลลัพธ์การทำนายจะไม่ดี

สำหรับแนวทางการวิจัยในอนาคต สามารถสรุปได้ดังนี้

(1) การพัฒนาเทคนิคคอมมูนิตี้ไมนิ่งซึ่งมีการใช้ทั้งข้อมูลเกี่ยวกับการเชื่อมต่อกันระหว่างเว็บเพจ และข้อมูลที่เป็นเนื้อความของเว็บเพจ

(2) การพัฒนาเว็บครอลเลอร์ที่มีความสามารถในการทำคอมมูนิตี้ไมนิ่ง

(3) การสร้างเว็บอาร์ไคว์ภาษาไทย เพื่อรองรับงานวิจัยด้านเว็บไมนิ่ง

บรรณานุกรม

Albert, R., Jeong, H., & Barabasi, A. (1999). The diameter of the world wide web. *Nature, 401(130).*

Anderson, R., & Lang, K. (2006). Communities from Seed Sets. In *Proceedings of the 15th International Conference on World Wide Web (WWW'06).*

Barabasi, A., & Albert R. (1999). Albert. Emergence of scaling in random networks. *Science,* 286.

Bernardo, A. H., Peter, P., James, E. P., & Rajan, M. L. (1998). Strong regularities in World Wide Web surfing. *Science, 280(5360),* 95–97.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3,* 993-1022.

Boldi, P., Codenotti, B., Santini, M., & Vigna, S. (2002). Structural properties of the African Web. In *Poster Proceedings of the 11th International Conference on World Wide Web (WWW'02).*

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Wiener, J. (2000). Graph structure in the Web. In *Proceedings of the 9th International Conference on World Wide Web (WWW'00).*

Cai, D., Mei, Q., Han, J., & Zhai, C. (2008). Modeling hidden topics on document manifold. In *Proceedings of ACM 17th Conference on Information and Knowledge Management (CIKM'08).*

Caminero, R. C., & Mikami, Y. (2008). The Link Structure of Language Communities and its Implication for Language-specific Crawling. In *The 6th Workshop on Asian Language Resources.*

Chakrabarti, S., Dom, B., Gibson, D., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1998). Experiments in topic distillation. In *SIGIR Workshop on Hypertext Information Retrieval on the Web.*

Chakrabarti, S., Berg, M., & Dom, B. (1999). Focused Crawling: a new approach to topic-specific Web resource discovery. In Proceedings of the 8th International World Wide Web Conference, 1623–1640.

Chang, J., & Blei, D. (2009). Relational topic models for document networks. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS'09).*

Chung, F. R. K. (1997). Spectral graph theory. *AMS Bookstore.*

Cohn, D., & Chang, H. (2000). Learning to probabilistically identify authoritative documents. In *Proceedings of the International Conference on Machine Learning (ICML'00)*

Cohn, D., & Hofmann, T. (2001). The missing link - a probabilistic model of document content and hypertext connectivity. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS'01).*

Davidson, B., D. (2000). Topical locality in the Web. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'00).*

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. B 39*, 1-38.

Dhillon, I., Guan, Y., & Kullis, B. (2007). Weighted graph cuts without eigen vectors : A Multilevel Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1944-1957.

Dhillon, I. S., Mallela, S., & Modha, D. S. (2003). Information-theoretic co-clustering. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03).*

Dietz, L., Bickel, S., Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the International Conference on Machine Learning (ICML'07).*

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed membership models of scientific publications. In *Proceedings of the National Academy of Science, 101*, 5220–5227.

Flake, G. W., Lawrence, S., & Gile, C. L. (2000). Efficient identification of web communities. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'00).*

Flake, G. W., Lawrence, S., Gile, C. L., & Coetzee, F. (2002) Self-organization of the Web and identification of communities. *IEEE Computer, 35(3),* 66-71.

Ford, L., & Fulkerson, D. (1956). Maximal flow through a network. *Canadian Journal of Mathematics, 8(3),* 399-404.

Garey, M. R., & Johnson, D. S. (1979). Computers and intractability: A guide to the theory of NP-completeness. W. H. Freeman, New York.

Gibson, D., Kleinberg, J., & Raghavan, P. (1998). Inferring Web communities from link topology. In *Proceedings of 9th ACM Conference on Hypertext and Hypermedia.*

Gibson, D., Kumar, R., McCurley, K. S., & Tomkins, A. (2006). Dense subgraph extraction. In *Mining Graph Data*, 411-441.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society.*

Griffiths, T. L., & Steyvers, M. (2003). Prediction and semantic association. In *Neural Information Processing Systems 15.* Cambridge, MA: MIT Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. In *Proceedings of the National Academy of Science, 101,* 5228-5235.

Gruber, A., Rosen-Zvi, M., & Weiss, Y. (2008). Latent topic models for hypertext. In *Proceedings of the 24th conference on Uncertainty in artificial intelligence (UAI'08)*

Heinrich, G. (2009) Parameter estimation for text analysis. *Technical Report, Fraunhofer IGD, Darmstadt, Germany.*

Hoffman, M., Blei, D., & Bach, F. (2010). On-line learning for latent dirichlet allocation. In *Neural Information Processing Systems.*

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence.*

Hofmann, T., Puzicha, J., & Jordan, M. I. (1999). Unsupervised learning from dyadic data. In *Advances in Neural Information Processing Systems,* volume 11. MIT Press.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic latent semantic analysis. *Machine Learning Journal, 42(1),* 177-196.

Ino, H., Kudo, M., & Nakamura, A. (2005). Partitioning of Web graphs by community topology. In *Proceedings of the 14th International Conference on World Wide Web (WWW '05).*

Karypis, G., & Kumar, V. (1999). Parallel multilevel k-way partitioning for irregular graphs. *SIAM Review, 41(2):278-300.*

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms.*

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the Web for emerging cyber-communities. In *Proceedings of the 8th International Conference on World Wide Web.*

Langville, A. N., & Meyer, C. D. (2006). Google's PageRank and beyond: The Science of Search Engine Rankings. *Princeton University Press.*

Leskovec, J., Lang, K. J., & Mahoney, M. W. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10).*

Li, L., Otsuka, S., & Kitsuregawa M. (2010). Finding related search engine queries by Web community based query enrichment. *World Wide Web, 13*(1-2), 121-142.

Liu, B. (2007). Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. *Springer*.

Mei, Q., Deng, C., Zhang, D., & Zhai, C. (2008). Topic modeling with network regularization. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*

Nallapati, R. M., Ahmed, A., Xing, E. P., & Cohen, W. W. (2008). Joint latent topic models for text and citations. In *Proceedings of The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08).*

Otsuka, S., Toyoda, M., Hirai, J., & Kitsuregawa, M. (2004). Extracting user behavior by Web communities technology on global Web logs. In *Proceedings of the 15th International Conference on Database and Expert Systems Applications (DEXA'04).*

Pierrakos, D., Paliouras, G., Papatheodorou, C., Karkaletsis, V., & Dikaiakos, M. D. (2003). Web community directories: A new approach to Web personalization. In *Proceedings of the 1st European Web Mining Forum (EWMF'03).*

Somboonviwat, K., Tamura, T., & Kitsuregawa, M. (2006). Simulation Study of Language Specific Web Crawling. In *Proceedings of the 21st International Conference on Data Engineering Workshops (ICDEW'05).*

Somboonviwat, K. (2008). Research on language specific crawling and building of Thai Web archive. *Unpublished doctoral dissertation, University of Tokyo.*

Somboonviwat, K., Suzuki, S., & Kitsuregawa, M. (2008). Connectivity of the Thai Web Graph. In *Proceedings of the 10th Asia-Pacific Web Conference on Progress in WWW research and development, Springer, 613–624.*

Sun, Y., Han, J., Gao, J., & Yu, Y. (2009). iTopicModel: information network-integrated topic modeling. In *Proceedings of 2009 International Conference on Data Mining (ICDM'09).*

Tamura, T., Somboonviwat, K., & Kitsuregawa, M. (2007). A method for language-specific web crawling and its evaluation. *Systems and Computers in Japan, 38(2),* 10–20.

Toyoda, M., & Kitsuregawa, M. (2001). Creating a Web community chart for navigating related communities. In *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia (HT'01).*

Toyoda, M., & Kitsuregawa, M. (2003). Extracting evolution of Web communities from a series of Web archives. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia (HT'03).*

Yang, T., Jin, R., Chi, Y., and Zhu, S. (2009). Combining link and content for community detection: a discriminative approach. In *Proceedings of The 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09).*

Yang, T., Chi, Y., Zhu, S., and Jin, R. (2010). Directed network community detection: A popularity and productivity link model. In *Proceedings of the 2010 SIAM International Conference On Data Mining (SDM'10).*

Yu, S., Moor, B. D., & Moreau, Y. (2009). Clustering by heterogeneous data fusion: framework and applications. In *NIPS workshop*.

Zhang, Y., Xu Yu, J., & Hou, J. (2006). *Web communities: analysis and construction.* Springer.

Zhu, S., Yu, K., Chi, Y., Gong, Y. (2007). Combining content and link for classification using matrix factorization. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'07).*

ภาคผนวก ก

ผลงานวิจัยที่เกี่ยวข้องกับการทำโครงการวิจัยและได้รับการตีพิมพ์เผยแพร่

- Somboonviwat, K. (2012). Web community analysis and its application to language specific crawling. In *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication* (ICUIMC '12). Article 98, 5 pages. DOI=10.1145/2184751.2184865

- Somboonviwat, K. (2013). Topic Modeling for Web Community Discovery. In G. Xu, & L. Li (Eds.), *Social Media Mining and Social Network Analysis: Emerging Research* (pp. 72-89). Hershey, PA: Information Science Reference. doi:10.4018/978-1-4666-2806-9.ch005

Web Community Analysis and its Application to Language Specific Crawling

Kulwadee Somboonviwat
International Software Engineering Program, International College
King Mongkut's Institute of Technology Ladkrabang (KMITL)
Chalongkrung Road, Ladkrabang
Bangkok 10520 Thailand
kskulwad@kmitl.ac.th

ABSTRACT

This paper proposes a novel metric for web community analysis, called *language homogeneity*. The language homogeneity of a community measures the ratio of web pages in a specific language within the community. This simple web community analysis can provide additional insights on the characteristics of web communities. We analyze web communities extracted from large Thai web datasets in the following aspects: (1) community size distribution, (2) similarity with a web directory, and (3) Thai language homogeneity. Interestingly, we found that most Thai web communities are linguistically homogeneous. Web pages inside the same community tend to be written in the same language. Based on these analysis results, we argue that the linguistic homogeneity of web communities can be used to enhance language specific crawling. Towards this end, we point out current limitations of a language specific crawler and suggest possible ways for exploiting communities' language homogeneity to improve the performance of language specific crawling.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *information networks*. H.2.8 [**Database Management**]: Database Applications – *data mining*.

General Terms

Measurement, Experimentation.

Keywords

web community mining and analysis, language homogeneity, language specific web crawling.

1. INTRODUCTION

The explosive growth of the World Wide Web has led to the excessive abundance of information. Many techniques have been developed to cope with this information explosion phenomenon.

In the early day of the Web, a web directory was devise as a tool to facilitate information search on the Web. Because a web directory based approach relies on human efforts to manually develop hierarchical listings of hyperlinks to web pages on a wide range of general topics. For this reason, the web directories might not be able to (1) provide the most up-to-date information on some topics, (2) detect emerging topics, and (3) provide information about very specialized topics.

To cope with these limitations, Kumar et al. [6] proposed a method to automatically extract implicit web communities from a large web crawl dataset. A web community is a set of web pages with coherence content and dense link structure. Implicit web communities provide invaluable, reliable, timely, and up-to-date topic specific information resources for users interested in them. Furthermore, a set of extracted web communities can be used as a key building block in other web applications and value added services such as focused crawling, web portals, related-pages search, web spamming detection, web recommendation, and web personalization.

This paper investigates the possibility of applying web community analysis to language specific crawling. Specifically, we propose a novel metric for web community analysis, called *language homogeneity*. Language homogeneity of a community measures the ratio of web pages in a specific language within the community. As will be shown later, this simple web community analysis can provide additional insights on the characteristics of web communities. We analyzed web communities extracted from large Thai web datasets in the following aspects: (1) community size distribution, (2) similarity with a web directory, and (3) Thai language homogeneity. Interestingly, we found that most Thai web communities are linguistically homogeneous. Web pages inside the same community tend to be written in the same language. Based on these analysis results, we argue that the linguistic homogeneity of web communities can be used to enhance the performance of a language specific crawler. Towards this end, we point out current limitations of a language specific crawler and suggest possible ways for exploiting communities' language homogeneity to improve the performance of language specific crawling.

The paper is organized as follows. Section 2 describes the basic concepts of web community and language specific crawling. Section 3 presents our Thai web community analysis. Section 4 discusses the application of language homogeneity analysis to improving crawler's performance. Section 5 reviews related work. Section 6 concludes the paper.

2. BACKGROUND

2.1 Web Community

A *web community* is a set of web pages created by individuals or organizations with common interest on a specific topic. The web community usually manifests itself as a densely connected set of web pages with coherent topical content. Formally, the web community and the web community discovery task can be defined as follows.

Definition 1 (Webgraph) A *webgraph* is a directed graph $G(V,E)$ where V is a set of vertices (or nodes) corresponding to web pages, and E is a set of ordered pairs (u, v) corresponding to hyperlinks from page u to page v.

Definition 2 (Web Community) Given a finite set of $P = \{p_1, p_2, \ldots, p_n\}$ of web pages, a *community* is a pair $C=(T, M)$, where T is the community topic and $M \subseteq P$ is the set of all pages in P that shares the topic T. If $p_i \in M$, then p_i is said to be a member of the community C.

Definition 3 (Web Community Discovery) Given a collection of web pages P and a web graph G induced from P, the *Web Community Discovery task* is to extract k web communities $\{C_1, \ldots, C_k\}$, such that each C_i is a subgraph in G with dense connections and coherent content.

There have been several web community discovery methods proposed in recent years (e.g. [4,5,6,3,10]). These methods identify a community based on its link structure characteristics such as bipartite or densely connected subgraphs. In this paper, we extracted Thai web communities from our datasets using the method proposed in [10].

2.2 Language Specific Crawling

Language specific crawling ([7,9]) is a method for selectively harvesting web pages written in a specific language. The task of a language specific crawler can be defined as follows.

Definition 4 (Language Specific Crawling) Given a set of start seed URLs and a target language L^*, crawl regions of the Web pertaining to the language L^*, avoid downloading of pages written in languages other than L^* as much as possible.

The framework for a language specific crawler is based on the framework of focused crawling proposed by Chakrabarti et al. in [2]. A block diagram of a language specific crawler is shown in Figure 1. The key idea is to use a language classifier to guess the language of a downloaded web page, and use the linguistic relationships between connected web pages to guide the crawler. The datasets used in this paper were constructed using a language specific crawler as proposed in [7,9].

3. WEB COMMUNITY ANALYSIS

In this section, we present our web community analysis conducted on Thai web datasets and report the results. First, in Section 3.1, we will define our concept of language homogeneity of a web community and other related measures. Then, the Thai data sets are presented in Section 3.2. Finally, we report our web community analysis results in Section 3.3.

3.1 Method

Our experiments measure two important properties of web communities: *topical similarity with topics in a web directory (e.g. Yahoo!)*, and *language homogeneity*.

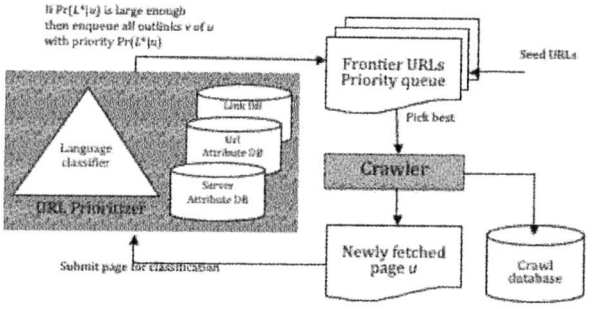


Figure 1. Block diagram of a language specific crawler.

Definition 5 (Topical Similarity) Let $C=(T, M)$ be a web community on topic T, whose members consist of web pages in set $M=\{m_1, m_2, \ldots, m_n\}$. Further, let $TD=\{s_1, s_2, \ldots, s_p\}$ be a topic directory with p categories and for each category s_j, j=1..p, there exists a set of web pages $U_{sj}=\{u_{sj,1}, u_{sj,2}, \ldots, u_{sj,q}\}$ belonging to that category. Then, the *topical similarity* (abbr. *TS*) between a community C and a category s_j in topic hierarchy TD is defined by

$$TS(C,s_j) = \frac{|M \cap U_{s_j}|}{|M|} \quad (1)$$

Here, $|M|$ is the number of member web pages in the community, $|M \cap U_{sj}|$ the number of elements of the set derived by intersecting M and U_{sj}, and $0 \le TS(C, s_j) \le 1$.

Definition 6 (Language Homogeneity) Let $C=(T, M)$ be a web community on topic T, whose members consist of web pages in set $M=\{m_1, m_2, \ldots, m_n\}$. Let l be a language and $M_l \subseteq M$ be a set of web pages in M that are written in language l. Then, language homogeneity (abbr. *LH*) of community C with respect to language l is defined by

$$LH_l = \frac{|M_l|}{|M|} \quad (2)$$

Here, $|M_l|$ is the number of pages in the community that are written in language l, and $0 \le LH_l \le 1$.

3.2 Thai Web Datasets

The Thai web datasets used in this paper were crawled using a general purpose BFS crawler and a language specific crawler. The properties of each dataset are shown in Table 1. For details on the criteria used in our language specific crawler, please refer to [8].

3.3 Analysis Results

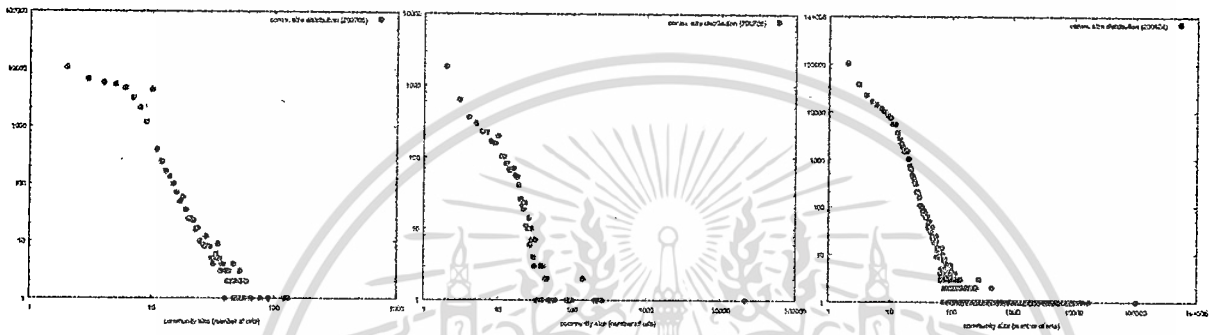
Applying the Companion- algorithm proposed in [10], we can extract a large number of web communities from our Thai web datasets. The number of communities found in each dataset is reported in Table 1.

3.3.1 Community size distribution

The log-log plots of size distributions of web communities extracted from the 200705, 200708, and 200804 datasets are as shown in Figure 2. As can be seen from Figure 2, the community size distributions of all datasets fit with the power-law distribution, which is ubiquitous in the Web graph [9]. Most communities are small with members less than 10 pages.

Table 1. The Thai web datasets. LS refers to language specific crawling and BFS refers to Breadth First Search crawling methods.

Dataset:	200705	200706	200707	200708	200804
Crawling method	LS	LS	LS	LS	BFS
HTML pages	1,188,217	2,063,858	2,451,383	2,295,208	68,839,605
Thai language HTML pages	502,909	420,074	635,450	602,528	3,633,975
Ratio of Thai pages in the dataset	0.42	0.20	0.25	0.26	0.05
Number of communities	45,443	6,199	6,781	4,977	265,476



(a) 200705 (b) 200708 (c) 200804

Figure 2. Community size distributions.

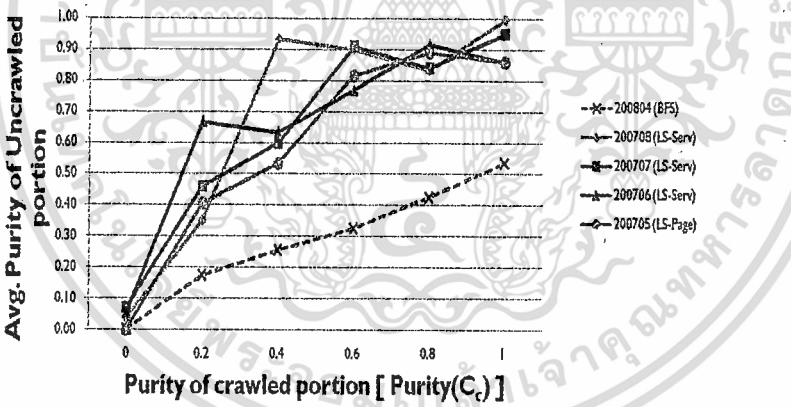


Figure 3. Correlations of the average language homogeneity (or purity) of the crawled and uncrawled portions.

Table 2. Thai pages in the uncrawled portions of web communities with $LH_{Thai} > 0.5$.

Dataset:	200705	200706	200707	200708	200804
Number of uncrawled pages	155	233	201	170	439
Number of uncrawled Thai pages	126	194	183	164	165
Ratio of Thai pages in the uncrawled portion	0.81	0.83	0.91	0.96	0.38

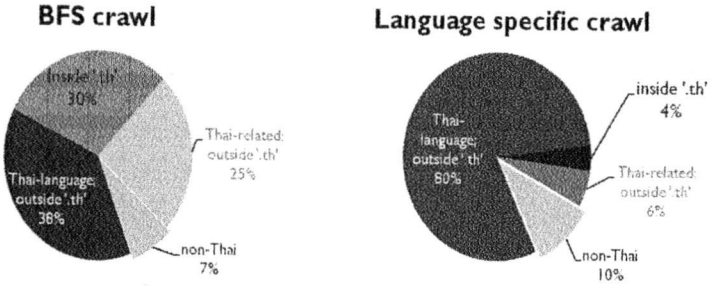


Figure 4. Types of web pages inside the uncrawled portions of communities with high language homogeneity values.

However, there are also a few extremely large web communities. By manual inspection, we found that most of those huge communities are blog web sites, spams, or pornographic websites. This explains why our language specific crawler achieved lower ratio of Thai pages on the 200708 dataset, the crawler was bogged down in communities of malicious content.

3.3.2 Topical Similarity and Language Homogeneity

To measure the topical similarity, we compared Thai communities with a Thai web directory available at http://www.truehits.net/. At the time of our experiment, this web directory consists of about 10,000 URLs classified into 19 different categories. We sampled some web communities and calculated the topical similarity *TS* defined in Definition 5. The result indicated high similarity between the web communities and the web directory's categories with some pairs of them reaching the *TS* value as high as 0.9.

Regarding the language homogeneity of web communities, we found that generally web communities extracted from language specific crawling datasets will be more homogenous than those extracted from breath-first crawling datasets.

To better understand, the language homogeneity of the Thai web communities (LH_{Thai}). We conduct another experiment. This time we divided the set of URLs within a community into two sets. The first set of URLs corresponds to URLs that have already been downloaded by the crawler. We will refer to the first set as the *crawled portion of a community* (C_c). The second set of URLs corresponds to URLs that have *not* been downloaded yet. We will refer to the second set as the *uncrawled portion of a community* (C_u). Then, for each portion of each community, we calculated its LH_{Thai} values and studied the correlations between the *LH* values of the crawled and uncrawled portions. The results of the experiment are as shown in Figure 3 and Table 2.

According to Figure 3 and Table 2, our observations are as follows.

- Web communities extracted from the language specific crawling datasets are more linguistically homogenous. This result shows that the language specific crawler is superior to the BFS crawler in collecting language specific web pages.
- There is a strong correlation between the LH_{Thai} of the crawled and uncrawled portions of a community. If the LH_{Thai} of the crawled portion is high, then there is a high probability that the LH_{Thai} of the uncrawled portion will also be high (>0.5).
- Communities $LH_{Thai} > 0.5$ may be used as a source to provide more URLs of Thai web pages to the crawlers.

4. Applying Web Community Analysis to Language Specific Crawling

In this section, we discuss our idea on applying the language homogeneity analysis to enhance the performance of a language specific crawler. The main idea is to guess the languages of the uncrawled web pages in a web community based on the language homogeneity values of web pages in the crawled portion.

According to our experiment in the previous section, if most of the crawled pages in a community are Thai related web pages, then most of the URLs in the uncrawled portion will be likely to link to Thai related web resources. Therefore, we can enhance the crawl seed set (i.e. the starting set of URLs given to the crawler) by including uncrawled URLs of the communities with high average language homogeneity value in the crawled portion (that is the communities C having C_c with $LH_{Thai} > 0.5$).

In order to show the feasibility of applying this concept to improve the performance of Thai web crawling, we conducted yet another experiment to study what kinds of web pages are included in the uncrawled portions of the communities with $LH_{Thai} > 0.5$. Figure 4 shows the classification of web pages types derived from the uncrawled portions of the communities with high value of LH_{Thai}. It can be seen that 90% of URLs in the un-crawled portion are Thai-related web pages. In addition, many Thai-related foreign language pages can be found from the uncrawled portions of these highly linguistic homogeneous communities.

To sum up, our results suggest that it would be possible and highly beneficial to apply the language homogeneity analysis for language specific crawling performance improvement, especially as a way to provide good starting seed URLs for the crawler.

5. RELATED WORK

In 1998, Kleinberg [5] proposed HITS (Hypertext Induced Topic Search) algorithm. The HITS algorithm is based on the co-citation relationship between two types of web pages, called *hubs* and *authorities*. An authority is a page with authoritative content on a topic and is linked to by many pages. A hub is a page with many links to good authority pages on a topic. Given a user query, HITS returns a ranked list of hubs and a ranked list of authorizes as query results. The computation of these two ranked lists is defined based on the mutual reinforcement relationship between hubs and authority. Later, Gibson et al. [4] showed that the HITS algorithm can be used to identify some communities. Based on [4], Kumar et al. [6] argued that a web community contains a bipartite

subgraph as its core and developed an efficient method for enumerating bipartite cores from a large web graph. Flake et al. [3] defined a web community as a subgraph whose internal hyperlinking is denser than hyperlinking to external web pages. [3] casted the web community discovery problem as a max-flow/min-cut problem and proposed an efficient method for identifying web communities. Toyoda et al. [10] introduced a method to constructing a web community chart that can reveal relationships among web communities.

Language specific web crawling was first studied by Tamura et al. [7] and Somboonviwat et al. [9] as a tool for constructing large web archives for a country. In [9], a crawler simulator was used to study linguistic characteristics of the Thai Web. Based on these linguistic characteristics and some additional graphical analysis of the Thai Web space, [7] proposed an efficient method for language specific web crawling.

Caminero et al. [1] analyzed the link structure of web pages under the South East Asian ccLTDs (country code top-level domains). They reported linguistic relationships between linked pages, size and diameter of the strongly connected components (SCC) of web pages written in the same language. Note that, Caminero et al. [1] did not consider the web pages beyond the country's top-level domain, which is an important challenge for a language specific crawler. Furthermore, they did not analyze implicit web communities that might also be very effective to enhancing the performance of a language specific crawler.

6. CONCLUSION

In this paper, we have put forward the idea of applying web community analysis to improving the performance of language specific web crawling. We also proposed a novel metric for web community analysis, called language homogeneity. We applied our web analysis methods to large Thai web datasets. The results suggested possibilities of using the proposed web community analysis in real web crawling.

For the future work, we plan to improve the performance of a real language specific crawler by intelligently exploiting the concept of language homogeneity in real-world language specific web crawling.

7. REFERENCES

[1] Caminero, R. C., & Mikami, Y. (2008). The Link Structure of Language Communities and its Implication for Language-specific Crawling. In *The 6th Workshop on Asian Language Resources*.

[2] Chakrabarti, S., Berg, M., & Dom, B. (1999). Focused Crawling: a new approach to topic-specific Web resource discovery. In *Proceedings of the 8th International World Wide Web Conference*, 1623–1640.

[3] Flake, G. W., Lawrence, S., & Gile, C. L. (2000). Efficient Identification of Web Communities. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[4] Gibson, D., Kleinberg, J., & Raghavan, P. (1998). Inferring Web Communities from Link Topology. In *Proceedings of 9th ACM Conference on Hypertext and Hypermedia*.

[5] Kleinberg, J. (1998). Authoritative Sources in a Hyperlinked Environment. In *Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms*.

[6] Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999).Trawling the Web for Emerging Cyber-Communities. In *Proceedings of the 8th International Conference on World Wide Web*.

[7] Tamura, T., Somboonviwat, K., & Kitsuregawa, M. (2007). A method for language-specific web crawling and its evaluation. *Systems and Computers in Japan, 38(2)*, 10–20.

[8] Somboonviwat, K., Suzuki, S., & Kitsuregawa, M. (2008). Connectivity of the Thai Web Graph. In *Proceedings of the 10th Asia-Pacific Web Conference on Progress in WWW research and development, Springer, 613–624*.

[9] Somboonviwat, K., Tamura, T., & Kitsuregawa, M. (2006). Simulation Study of Language Specific Web Crawling. In *Proceedings of the 21st International Conference on Data Engineering Workshops (ICDEW'05)*.

[10] Toyoda, M., & Kitsuregawa, M. (2001). Creating a Web Community Chart for Navigating Related Communities. In *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia (HT'01)*.

Social Media Mining and Social Network Analysis:

Emerging Research

Guandong Xu
University of Technology Sydney, Australia

Lin Li
Wuhan University of Technology, China

Information Science
REFERENCE

Managing Director:	Lindsay Johnston
Editorial Director:	Joel Gamon
Book Production Manager:	Jennifer Yoder
Publishing Systems Analyst:	Adrienne Freeland
Development Editor:	Monica Speca
Assistant Acquisitions Editor:	Kayla Wolfe
Typesetter:	Christy Fic
Cover Design:	Nick Newcomer

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Social media mining and social network analysis: emerging research / Guandong Xu and Lin Li, editors.
p. cm.
Includes bibliographical references and index.
Summary: "This book highlights the advancements made in social network analysis and social web mining and their influence in the fields of computer science, information systems, sociology, organization science discipline and much more"--Provided by publisher.
ISBN 978-1-4666-2806-9 (hbk.) -- ISBN 978-1-4666-2807-6 (ebook) -- ISBN 978-1-4666-2808-3 (print & perpetual access) 1. Online social networks--Research. 2. Data mining. 3. Communication--Network analysis. I. Xu, Guandong. II. Li, Lin, 1977-
HM742.S6282 2013
006.7'54--dc23

2012033295

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 5
Topic Modeling for Web Community Discovery

Kulwadee Somboonviwat
King Mongkut's Institute of Technology Ladkrabang (KMITL), Thailand

ABSTRACT

The proliferation of the Web has led to the simultaneous explosive growth of both textual and link information. Many techniques have been developed to cope with this information explosion phenomenon. Early efforts include the development of non-Bayesian Web community discovery methods that exploit only link information to identify groups of topical coherent Web pages. Most non-Bayesian methods produce hard clustering results and cannot provide semantic interpretation. Recently, there has been growing interest in applying Bayesian-based approaches to discovering Web community. The Bayesian approaches for Web community discovery possess many good characteristics such as soft clustering results and ability to provide semantic interpretation of the extracted communities. This chapter presents a systematic survey and discussions of non-Bayesian and Bayesian-based approaches to the Web community discovery problem.

INTRODUCTION

In recent years, the World Wide Web has become a popular platform for disseminating and searching for information. Due to the explosive growth of the Web, the low precision of Web search engine, and the lack of a data model for the Web data, it is increasingly difficult for the users to search for and access the desired information. Motivated by this problem, a lot of research has been done to discover the implicit communities of topically related Web pages or *Web communities* (e.g. Gibson, et al., 1998; Kumar, et al., 1999; Flake, et al., 2000). The Web communities provide invaluable, reliable, timely, and up-to-date topic specific information resources for users interested in them. Furthermore, a set of extracted Web communities can be used as a key building block in Web applications and value added services such as focused crawling, Web portals, Web search ranking, Web spamming detection, Web recommendation, and Web personalization (such as Flake, et al., 2000;

DOI: 10.4018/978-1-4666-2806-9.ch005

Pierrakos, et al., 2003; Otsuka, et al., 2004; Li, et al., 2010).

Conceptually, a Web community is defined as a set of Web pages on a specific topic created by people sharing the same interests. The Web community usually manifests itself as a subgraph with dense connections and coherent content. Most work on Web community discovery focused on the efficient detection of community structure based purely on link information between Web pages using non-Bayesian approaches such as spectral methods, graph partitioning, and clustering (e.g. Kumar, et al., 1999; Flake, et al., 2000; Toyoda & Kitsuregawa, 2001). These non-Bayesian link based methods suffer from the lack of semantic interpretation (most implementation uses a simple top-k most frequent keyword to summarize a topic of a Web community). Furthermore, most non-Bayesian approaches for Web community discovery do not allow a Web page to be assigned to more than one community.

On the other hand, probabilistic topic models (e.g. Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2002, 2003, 2004; Hofmann, 1999, 2001) have recently gained much popularity as a suite of algorithmic tools to help organizing, searching, and understanding large collections of text documents. The key idea underlying these models is that a document is generated from a probabilistic process, and consists of multiple topics, where a topic is a probability distribution over words taken from a fixed set of vocabularies. This representation naturally captures the hidden topical structure in text, and can then be used in text mining tasks to discover topics from a large text collection.

With the explosive growth of the Web and linked data sets, some recent work on topic modeling has extended the basic topic models by taking into consideration the link structure information. The work in this area can be classified into five directions. The first line of work (e.g. PHITS-PLSA by Cohn & Hofmann, 2001; LDA-Link-Word by Erosheva, et al., 2004; Link-PLSA-LDA by Nallapati, et al., 2008) incorporates the notion of link information into the document generative model. The second line of work (relational or supervised topic models) models textual content and link separately by representing the link between documents as a binary random variable conditioned on their content (e.g. Chang & Blei, 2009). The third line of work regularizes topic models with a discrete regularizer defined based on the link structure of the data set (e.g. NetPLSI by Mei, et al., 2008). The fourth line of work (e.g. iTopicModel by Sun, et al., 2009) model the relationship between documents using a multivariate Markov Random Field (MRF). Lastly, Yang et al. (2009) proposed the PCL model, which is a discriminative model for combining link and content information for community detection.

Probabilistic topic modeling possesses many characteristics that are desirable for the Web community discovery problem. Firstly, they produce soft clustering results therefore it is possible for a Web page to belong to multiple communities (i.e. they can produce overlapping community structure). Secondly, by leveraging the textual information, they are able to provide semantic interpretation of the extracted communities.

This chapter presents a systematic review and discussions of research efforts on Web community discovery problem. The plan of the chapter is as follows. First, we present background concepts related to Web community discovery and probabilistic topic models. Next, we formulate the Web community discovery problem, and present some representative non-Bayesian community discovery algorithms. Then, we describe key concepts underlying probabilistic topic models in details, and discuss some recent work on Bayesian based methods for Web community discovery. Finally, we highlight some interesting future research directions and conclude the chapter.

BACKGROUND

Graph Terminologies

Definition 1 (Directed graph): A *directed graph* *G(V,E)* consists of a set *V* of nodes (or vertices) and a set *E* of ordered pairs of nodes, called edges. Each edge is an ordered pair of nodes *(u,v)* representing a directed connection from node *u* to node *v*.

Definition 2 (In-degree and Out-degree): The *in-degree* of a node *u* is the number of distinct incoming edges incident to *u* (i.e. the number of distinct edges (v_1,u), (v_2,u), ..., (v_k, u). The *out-degree* of a node *u* is the number of distinct outgoing edges incident to *u* (i.e. the number of distinct edges (u,v_1), (u,v_2), ..., (u,v_k).

Definition 3 (Path and Distance): A *path* from node *u* to node *v* is a sequence of edges (u,u_1), (u_1,u_2), ... (u_k,v). The *length of the path* is the number of edges along the path. The *distance* from node *u* to node *v* is equal to the length of the smallest path from *u* to *v* for which such a path exists. If no path exists, the distance from *u* to *v* is defined to be infinity.

Definition 4 (Diameter): A *diameter* of a graph *G(V,E)* is the greatest *distance* between any pair of nodes *(u,v)* in *G*.

Definition 5 (Webgraph): A *webgraph* is a directed graph *G(V,E)* where *V* is a set of nodes corresponding to Web pages, and *E* is a set of ordered pairs *(u,v)* corresponding to hyperlinks from page *u* to page *v*.

Definition 6 (Directed bipartite graph): A *directed bipartite graph* is a directed graph *G(V,E)* whose node set can be partitioned as $V = F \cup C$, with the property that every edge $e \in E$ has one end in *F* and the other end in *C*.

Definition 7 (Complete directed bipartite graph): A *complete directed bipartite graph* is a directed bipartite graph $G_C(V,E)$ whose node set can be partitioned as $V = F \cup C$, such that every node in *F* has an edge to every node in *C*. A complete bipartite graph is denoted as $G_{Cf,c}$ where $f = |F|$ and $c = |C|$.

A complete directed bipartite graph, $G_{C4,4}$ is shown in Figure 1. The node set of the graph $G_{C4,4}$ can be divided into two disjoint sets $F = \{ f_1, f_2, f_3, f_4 \}$ *and* $C = \{ c_1, c_2, c_3, c_4 \}$, such that each node f_i (i=1,2,3,4) in *F* has edges to every node c_j (j=1,2,3,4) in *C*.

Definition 8 (Dense directed bipartite graph) A *dense directed bipartite graph* is a directed bipartite graph $G_D(V,E)$ whose node set can be partitioned as $V = F \cup C$, such that a node in *F* must has edges to at least γ_C ($1 \le \gamma_C \le |C|$) nodes in *C*, and at least γ_F ($1 \le \gamma_F \le |F|$) nodes in *F* link to every node in *C*. A dense directed bipartite graph is denoted as $G_D\gamma_C\gamma_F$.

A dense directed bipartite graph, $G_{D3,2}$ is shown in Figure 2. The node set of the graph $G_{D3,2}$ can be divided into two disjoint sets $F = \{ f_1, f_2, f_3, f_4 \}$ *and* $C = \{ c_1, c_2, c_3, c_4 \}$, such that a node f_i in *F* has edges to at least γ_C=3 in *C*, and at least γ_F=2 in *F* link to every node in *C*.

HITS Algorithm

HITS (Hypertext Induced Topic Search) (Kleinberg, 1998) is a search query dependent ranking algorithm that produces two rankings of Web

Figure 1. A complete directed bipartite graph $G_{C4,4}$

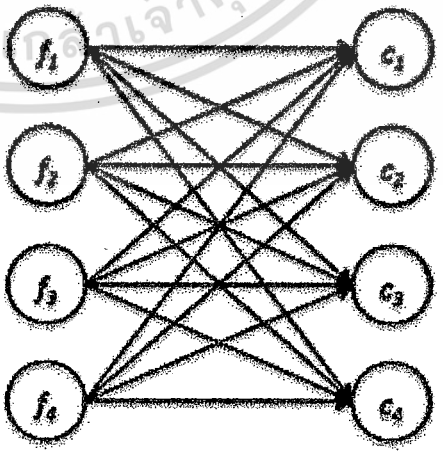
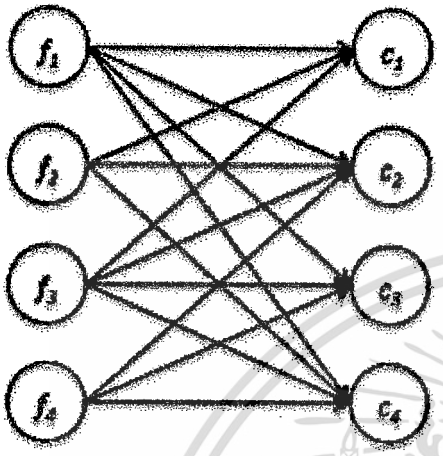


Figure 2. A dense directed bipartite graph $G_{D3,2}$



pages in response to a user query: an *authority ranking* and a *hub ranking*. An *authority* is a page with authoritative content on a topic and is linked to by many pages. A *hub* is a page with many links to good authority pages on a topic. The key idea underlying HITS is the concept of *mutual reinforcement relationship* between authority and hub pages (i.e. a good hub points to many good authorities, and a good authority is pointed to by many good hubs). Topologically, the mutual reinforcement relationship usually manifests itself as a set of *dense bipartite subgraph G(V,E); V=F∪C* where *F is a set of hub pages*, and *C is a set of authority pages*.

The HITS algorithm consists of two phases. Given an input query from a user, the first phase constructs a base set *S*, which is a focused Web subgraph containing many good authorities using Web search engines. The second phase computes an authority score and a hub score for every page in the base set *S*. Let *n* be the number of pages in *S*; *G(V,E)* denotes the hyperlink graph induced from *S*, and *L* is an adjacency matrix of the hyperlink graph *G(V,E)*. Further, let the authority score of page *i* be *a(i)* and the hub score of page *i* be *h(i)*. We define a column vector of hub scores $h=(h(1),...,h(n))^T$ and a column vector of authority scores $a=(a(1),...,a(n))^T$. Based on the mutual reinforcement relationship, the authority and the hub scores can be defined by

$$a = L^T h \tag{1}$$

$$h = La \tag{2}$$

The authority and hub scores in Equation (1) and (2) can be computed using an iterative algorithm (see Algorithm 1). Note that, the following equations in Algorithm 1

$$\begin{aligned} a_k &= L^T h_{k-1} \\ h_k &= L a_k \end{aligned} \tag{3}$$

can be simplified to

$$\begin{aligned} a_k &= L^T L a_{k-1} \\ h_k &= L L^T h_{k-1} \end{aligned} \tag{4}$$

The two equations in (4) define the *power iteration method* for computing the *principal eigenvector* for the matrices L^TL and LL^T. Therefore, the computation of the authority and hub scores boils down to finding the principal right-hand eigenvectors of L^TL (the authority matrix) and LL^T (the hub matrix), respectively. Readers interested in detailed discussions of the HITS algorithm are directed to Kleinberg, 1998; Liu, 2007; Langville & Meyer, 2006 (see Algorithm 1).

The Small-World Phenomenon

The *small-world phenomenon* of the Webgraph is described by the fact that the diameter of the Webgraph is on average small relative to the size of the overall graph (i.e. the diameter is bounded by a polynomial in log *n* where *n* is the number of nodes in the graph). The small-world phenomenon has been reported in many subgraphs of the Web spanning from a university, a country, and a global

Algorithm 1. HTS algorithm

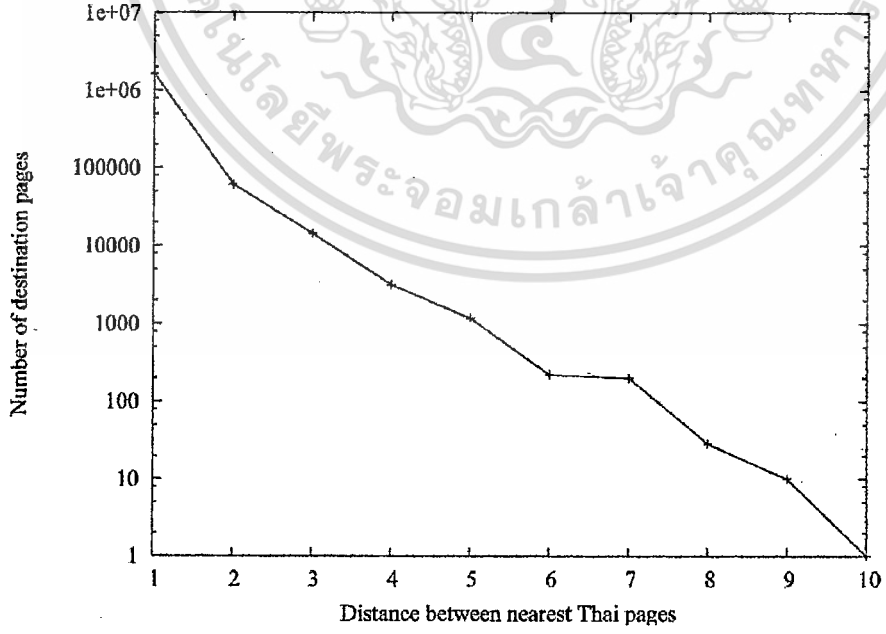
Let a_k, h_k denote authority and hub scores at the kth iteration
Initialize: $h_0 = (1,1,...,1)^T$
Until convergence, do
$a_k = L^T h_{k-1}$
$h_k = L a_k$
$k = k + 1$
Normalize a_k and h_k (e.g. by using the 1-norm):
$a_k = a_k / \| a_k \|_1$
$h_k = h_k / \| h_k \|_1$

Web (see, for example Albert, et al., 1999; Barabasi & Albert, 1999; Broder, et al., 2000; Boldi, et al., 2002; Tamura, et al., 2007; Somboonviwat, 2008).

Due to the small diameter of the Webgraph, it is straightforward that the distance between any two nodes in a Webgraph is rather short. An important implication of the existence of relatively short distances between nodes in the Webgraph is that there exist potentially many small densely connected clusters of Web pages in the Webgraph. As we shall see later, a densely connected cluster of Web pages is the identifying characteristics of a Web community.

To demonstrate the existence of rather small distances between same types (i.e. same topics or same languages) of nodes in a Webgraph, let us consider Figure 3 which is a plot of the distance between nearest *Thai Web pages* in a Webgraph determined from a 4 million pages crawl of the Thai Web in 2004 (see Tamura, et al., 2007; Somboonviwat, 2008 for more details). Note that, the *Thai Web pages* refer to the Web pages that, based on the classification by a language classifier, are likely to be written in the Thai language. According to Figure 3, most Thai pages are located near to the other Thai pages in the Webgraph. The maximum distance between any two Thai pages in the data sets is only 10, which is relatively small compared to the size of the Thai Webgraph induced from our Thai Web dataset (the graph consists of 9,953,318 nodes and 123,836,342 non-duplicated directed

Figure 3. Distribution of distances between nearest Thai pages



edges). This result suggests the occurrence of the small-world phenomenon in the Thai Webgraph. Therefore, we can anticipate the existence of many small dense clusters of Thai Web pages (i.e. the communities of Thai pages) in the Webgraph associated with the Thai dataset. In fact, according to a Web community extraction experiment conducted by Somboonviwat (2008), more than 67,000 communities were identified on this Thai Webgraph. The distribution of community sizes as reported in Somboonviwat (2008) is shown in Figure 4.

Probabilistic Graphical Models

Probabilistic graphical models are graphical representations of probability distributions. They provide a simple way to visualize the structure of a probabilistic model. In a probabilistic graphical model, directed graphs are used to express causal relationships between random variables. Each node in a graph represents a random variable (or a group of random variables); a shaded node represents an observed random variable. Each edge in a graph represents probabilistic relationships between random variables. A directed edge from node u to node v corresponds to a conditional distribution $p(v|u)$. A *plate notation* can be used to more compactly represent a replicated structure in the model. Figure 5(a) shows a graphical model corresponding to a joint distribution in Equation (5). The equivalent plate notation for this graph is shown in Figure 5(b). Note that, there are N observed random variables in this graph i.e. the shaded nodes $t_1, t_2, \ldots, t_{N-1}, t_N$

Figure 4. Distribution of community sizes extracted from a Thai Web crawl

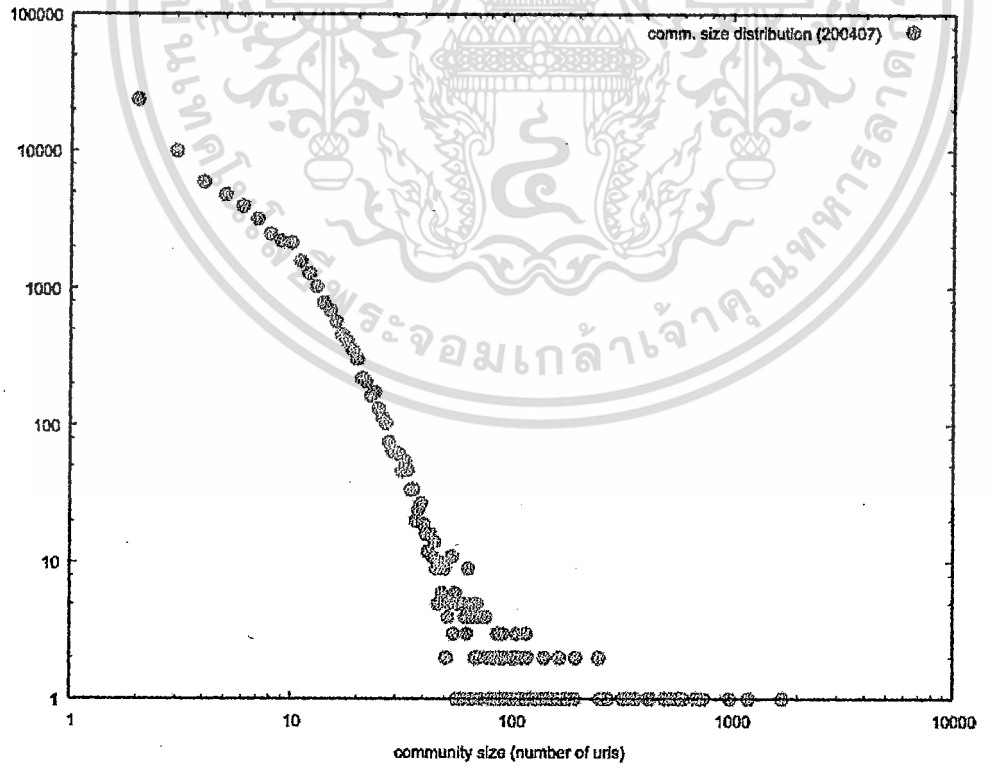
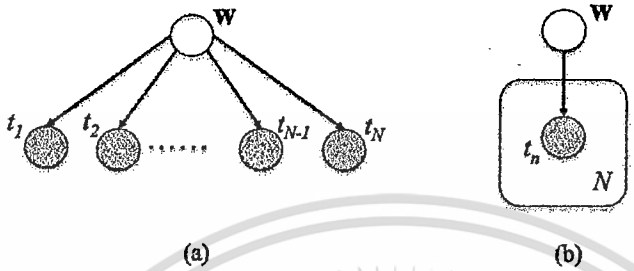


Figure 5. Graphical models: (a) directed graphical model representing the joint distribution in (5); (b) the same graphical model depicted using a plate notation



$$p(w, t_1, t_2, \ldots, t_{N-1}, t_N) = p(w)\prod_{n=1}^{N} p(t_n \mid w) \quad (5)$$

The Dirichlet Distribution

The *Dirichlet distribution* is an exponential family distribution that is a conjugate prior of the multinomial distribution. Given a multinomial distribution $\mu=(\mu_1, \ldots, \mu_K)$, such that $0 \le \mu_k \le 1$, $\sum_k \mu_k = 1$; and a parameter $\alpha=(\alpha_1, \ldots, \alpha_K)^T$. A K dimensional Dirichlet distribution for the multinomial μ is defined over a simplex of dimensionality $K-1$, and its normalized form is defined by

$$Dir(\mu \mid \alpha) = \frac{\Gamma(\sum_{k=1}^{K} a_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \mu_k^{\alpha_{k-1}} \quad (6)$$

Here $\Gamma(x)$ is a gamma function defined by

$$\Gamma(x) = \int_0^{\infty} u^{x-1} e^{-u} du \quad (7)$$

The parameter $\alpha=(\alpha_1, \ldots, \alpha_K)^T$ of the Dirichlet distribution controls the mean shape and the sparsity of the Dirichlet distribution as illustrated in Figure 6. For $\alpha < 1$ (see Figure 6[a]), the modes of the distribution are located at the corners of the simplex, leading to more sparse distribution of the density. For $\alpha > 1$ (see Figure 6[c]), the mode of the distribution is located away from the corners of the simplex, leading to more smooth distribution of the density.

Finally, we note that because the Dirichlet distribution is a conjugate prior of the multinomial distribution, therefore it follows that given a multinomial distribution μ, the posterior distribution of μ has the same functional form as the prior, i.e. the Dirichlet distribution. The conjugacy of the Dirichlet priors lead to greatly simplified Bayesian analysis.

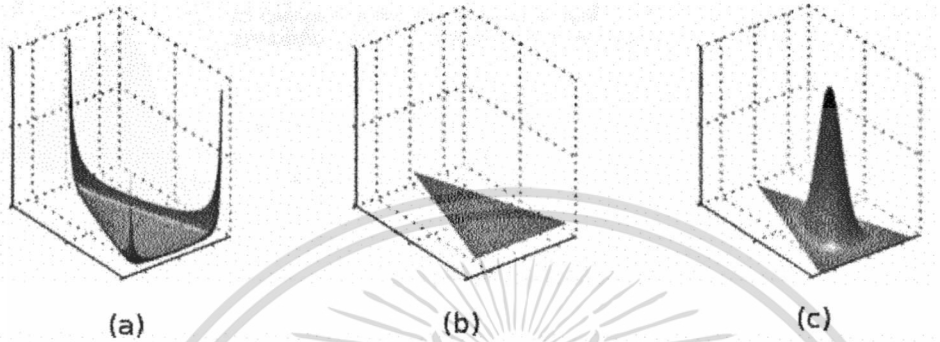
NON-BAYESIAN APPROACHES TO WEB COMMUNITY DISCOVERY

Problem Formulation

Conceptually, a *community* is a set of entities (e.g. Web pages, people, organizations) sharing a common interest. Based on a conceptual definition of a community given in Liu (2007), we define a *Web community* as follows.

Definition 9 (Web Community): Given a finite set of $P = \{p_1, p_2, \ldots, p_n\}$ of Web pages, a *community* is a pair $C=(T, M)$, where T is the *community topic* and $M \subseteq P$ is the set of all pages in P that shares the topic T. If $p_i \in M$, then p_i is said to be a *member of the community C*.

Figure 6. The three dimensional Dirichlet distribution. The two horizontal axes are coordinates in the plane of the simplex, and the vertical axis is the density. (a) $\alpha = (0.1,0.1,0.1)^T$, (b) $\alpha = (1,1,1)^T$, (c) $\alpha = (10,10,10)^T$.



Based on the concepts of Web communities defined above, we can formalize the task of *link based Web community discovery* as follows.

Definition 10 (Link Based Web Community Discovery): Given a collection of Web pages P and a Web graph G induced from P, the *Link Based Web Community Discovery task* is to extract k Web communities $\{C_1, ..., C_k\}$, where each C_i is a subgraph in G with dense connections and coherent content.

Definition 9 and 10 are conceptual definitions of a Web community and link based Web community discovery task. Different Web community discovery algorithms define their own operational definition of a Web community based on an assumption made regarding the specific forms of manifestation of Web communities in the dataset.

Bipartite Cores-Based Approach

As shown by Gibson et al. (1998), HITS can detect some Web communities based on broad topic queries using eigenvector computation. This result suggests that the footprint of a Web community can be recognized as a *dense bipartite subgraph* arises from the co-citation process, where related pages do not reference one another but are frequently referenced together. This could be caused by the fact that the Web sites are competitors or they do not share a common viewpoint, or simply because they are not aware of each other. Based on this observation, Kumar et al. (1999) argue that co-citation is a characteristic of well-developed explicitly known communities and is also an early indicator of emerging Web communities. Nevertheless, eigenvector computation as used by Gibson et al. (1998) is inefficient for iterating all bipartite Web communities. Kumar et al. (1999) proposed a heuristic based technique for automatically and efficiently enumerating all bipartite core communities from a large Web crawl. They used the term *trawling the Web* to denote this process. The notion of dense bipartite subgraphs representing Web communities can be formulated mathematically as follows.

Let $K_{ij} = G_C(V,E)$ be a *complete bipartite graph* whose nodes can be divided into two sets denoted F and C (where $V = F \cup C$; $|F| = i$, $|C| = j$), and every directed edge in E is directed from a node $f \in F$ to a node $c \in C$. Define an *(i, j)* core as a complete bipartite graph K_{ij} with optionally additional edges other than edges from F to C. According to this definition, an *(i, j) core* community contains a set of i pages all of which point to another set of j pages. Intuitively, the i pages are pages created by members of the community who wants

to link their pages to the most valuable resources (i.e. the *j* pages) for that community. Due to this reason, the *i* pages are called *fans*, and the *j* pages are called *centers*.

The trawling algorithm is based on two assumptions. The first assumption states that *any sufficiently strong Web community will be highly likely to contain an (i, j) core.* The second assumption states that *almost all occurrences of (i, j) cores are due to the existence of a community rather than random.* Based on these two assumptions, the main idea of the trawling algorithm is to identify a community by finding its core, and then to expand the core to the rest of the community. The trawling algorithm is presented in Algorithm 2. Note that dense subgraph extraction arises in many real-world graph analysis problems. Gibson et al. (2006) provides a detailed discussion on this problem, especially in the context of Web community discovery.

Maximum Flow-Based Approach

Flake et al. (2000) define a Web community based on density of links between Web pages. Given a Webgraph $G=(V, E)$, a community is a subset C of V such that each $v \in C$ has at least as many neighbors in C as in $V - C$. According to this definition, identifying a community is intractable because it maps into a family of NP-complete graph partitioning problems (Garey & Johnson, 1979). Flake et al. (2000) show that by assuming the existence of one or more seed websites and utilizing regularities of the Web graph (Bernado, et al., 1998; Barabasi, et al., 1999), a community can be identified by calculating the $s-t$ maximum flow of G and identifying nodes that are reachable from s to be the Web community.

The $s - t$ maximum flow problem is defined as follows. Given a directed graph $G=(V, E)$, with a source node $s \in V$, a target node $t \in V$, and edge capacities $c(u,v) > 0$, find the maximum flow that can be directed from the source node s to the target node t, without exceeding the capacity constraints on any edge. The Max Flow-Min Cut theorem of Ford and Fulkerson (1956) proves that the maximum flow of the network is identical to the capacity of the minimum cut that separates s and t. Many implementations exist for solving the maximum flow problem in polynomial time (Ravindra, et al., 1997). The Max Flow Web Community Discovery algorithm (Flake, et al., 2000, 2002) is illustrated in Algorithm 3. The procedure Max-Flow-Community (G, S, k) augments the

Algorithm 2. Trawling algorithm

STEP 1: Identifying the community cores *(i, j)*
STEP 1-1: Iterative pruning.
Until no node is qualified for deletion, do
Delete potential fans with out-degree less than *j*
Delete potential centers with in-degree less than *i*
STEP 1-2: Inclusion-exclusion pruning // Inclusion-exclusion based on fan pages
Until no further inclusion/exclusion, do
Choose a fan *u* with out-degree exactly *j*
Let $\Gamma(u)$ be a set of centers to which *u* points to
If there are *i-1* other fans all pointing to each center in $\Gamma(u)$
Include a new *(i, j)* core to the output
Otherwise
Exclude *u* from further contention (as a fan)
STEP 1-3: Exhaustive enumeration
Fix *j*, start with the *(1, j)* cores, one can construct all *(2, j)* cores by checking every fan that also points to any center in a *(1, j)* core. Similarly, all *(3, j)* can be constructed by checking every fan that points to any center in a *(2, j)* core. The procedure proceeds in a similar manner until the desired number of communities is extracted from the graph.
STEP 2: Core Expansion
Use algorithms, such as HITS (Kleinberg, 19998) or Clever (Chakrabarti et al., 1998), to expand the core.

input Webgraph G with an artificial source s and an artificial sink t. After augmenting the graph, a residual flow is generated by calling a maximum flow procedure Max-Flow (G, s, t). Then, all nodes reachable from s through non-zero positive edges form the output community. In comparison to the bipartite cores based approach, the max-flow based community discovery can extract larger, more complete communities. Like the bipartite-based approach, however, it cannot find the topic, and the relationships of Web communities.

Other Non-Bayesian Approaches

A great deal of work has been devoted to discovering communities in networks based on non-Bayesian methods. Gibson et al. (1998) show that the HITS algorithm can be used to extract some communities by computing the non-principal eigenvectors of the authority and hub matrices. Their results suggest that a Web community can manifest itself as a dense bipartite subgraph. Based on this intuition, Kumar et al. (1999) propose a heuristic based technique for finding all bipartite core communities efficiently from a large Web crawl dataset. Flake et al. (2000) cast the Web community discovery problem into the framework of the maximum flow model, and presented a maximum flow community discovery algorithm (see Algorithm 3). Toyoda and Kitsuregawa (2001, 2003) introduce a Web community chart that can identify relationships among Web communities, and present an algorithm to extracting relationship between Web communities and their evolution from a series of Japanese Web archives. Other recent Web community detection methods include e.g. Ino et al. (2005) and Anderson and Lang (2006).

Communities in networks may also be viewed as clusters, there are numerous methods from graph partitioning, clustering, and matrix reordering that have been applied to community detection. These include, for example, the METIS method (Karypis, et al., 1999), spectral analysis methods (Chung, 1997), multi-level clustering (Dhillon, et al., 2007), co-clustering (Dhillon, et al., 2007), matrix factorization (Zhu, et al., 2007), and Kernel fusion (Yu, et al., 2009). Readers interested in detailed discussions of Web communities analysis and construction are directed to Zhang et al. (2006).

Although link based community discovery has been very successful and applied in many real-world applications, a limitation exist when we not only want to know in an automated fashion the community memberships of each vertex in the network but also the semantics of the extracted

Algorithm 3. Max flow community discovery (Flake, et al., 2002)

Algorithm Find-Community
input: a set S of seed Web pages
while number of iteration is less than desired
do
Set $G = (V,E)$ to a fixed depth crawl from S.
Set $k = |S|$.
call: C =Max-Flow-Community(G, S, k)
Rank all $v \in C$ by number of edges in C.
Add the highest ranked non-seed nodes to S.
end while
output: all $v \in V$ still connected to the source s.

procedure Max-Flow-Community
input: graph $G=(V,E)$; a set $S \subset V$; integer k.
Create artificial nodes s and t, and add to V.
for all $v \in S$ **do**
Add (s,v) to E with $c(s,v)=\infty$.
end for
for all $(u,v) \in E$, $u \neq s$ **do**
Set $c(u, v) = k$.
if $(v, u) \notin E$ **then**
add (v, u) to E with $c(v, u) = k$.
end if
end for
for all $v \in V$, $v \notin S \cup \{s, t\}$ **do**
Add (v,t) to E with $c(v,t)=1$.
end for
call: Max-Flow(G, s, t).
output: all $v \in V$ still connected to s.

communities. One promising approach to providing the semantics in Web community extraction is the probabilistic topic modeling approach (e.g. Nallapati, et al., 2008; Chang & Blei, 2009; Mei, et al., 2008; Sun, et al., 2009; Yang, et al., 2009). The remaining of this chapter describes the basic concepts of topic modeling and recent efforts on the application of topic modeling to Web community discovery problem.

BAYESIAN APPROACHES TO WEB COMMUNITY DISCOVERY

Topic modeling has been successfully applied to discovering the topical structure of large text corpora. In this section, we first introduce two well-known traditional topic modeling methods, i.e. the Probabilistic Latent Semantic Indexing (PLSI) by Hoffman, 1999, and the Latent Dirichlet Allocation (LDA) by Blei et al. (2003). Then, we discuss several enhancements to the traditional topic modeling methods for Web community discovery.

Probabilistic Topic Modeling

Topic modeling (Hofmann, 1999, 2001; Blei, et al., 2003; Griffiths & Steyvers, 2002, 2003, 2004) is a suite of algorithms that extract a set of latent topics hidden inside a text corpus. The key idea of topic models is to represent a document in a latent space with a finite mixture model of k topics (i.e. a document can be viewed as consisting of multiple topics), where each topic is a probabilistic distribution over a set of words. The topic model defines a *generative model* for generating documents. The parameters in the generative model can be estimated by fitting the data with the model using posterior computation algorithms such as Gibbs sampling, and variational inference algorithms. Two well-known topic models are the Probabilistic Latent Semantic Analysis (PLSA) proposed by Hoffman (1999) and the Latent Dirichlet Allocation (LDA) proposed by Blei et al. (2003).

The PLSA model (Hofmann, 1999) is based on a latent variable model for co-occurrence data, called the aspect model (Hofmann, et al., 1999). The aspect model associates an unobserved latent class variable $z \in Z = \{z_1, ..., z_K\}$ with each occurrence of a word $w \in W = \{w_1, ..., w_M\}$ in a document $d \in D = \{d_1, ..., d_N\}$. A generative model of PLSA is defined as follows.

- Pick a document d with probability $P(d)$.
- Pick a latent topic z with probability $P(z|d)$.
- Generate a word w with probability $P(w|z)$.

The graphical model for PLSA is as shown in Figure 7(a). The generative process of the PLSA can be translated into a joint probability model as follows.

$$P(d,w) = P(d)P(w|d), where \tag{8}$$

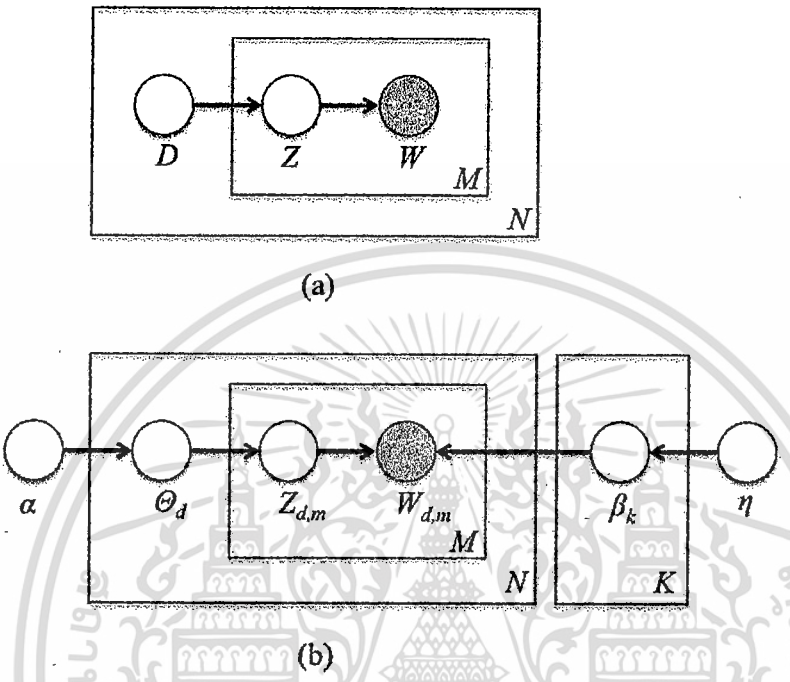
$$P(w|d) = \sum_{z \in Z} P(w|z)P(z|d) \tag{9}$$

Let $n(d,w)$ denotes the number of occurrences of the term w in a document d. The log likelihood of a document collection D to be generated is given by

$$L(D) = \sum_{d \in D} \sum_{w \in W} n(d,w) \log \sum_{j=1}^{K} p(w \mid z_j) p(z_j \mid d) \tag{10}$$

The posterior computation of PLSA can be done using the standard Expectation Maximization (EM) (Dempster, et al., 1977) algorithm. The EM algorithm alternates between an *expectation step* (E-step) and a maximization step (M-step). In the E-step, the posterior probabilities are calculated for the latent variables using the current estimates of the parameters. In the M-step, the model parameters are updated based on a local maximum

Figure 7. The graphical models for (a) PLSA, (b) LDA



of the log-likelihood in (10), which depends on the posterior calculated in the latest E-step. The posterior probabilities and the re-estimation equations for the E-step and the M-step for PLSA are given in Equation (11) - (14) below.

Expectation Step (E-step):

$$P(z \mid d,w) = \frac{P(z)P(d\mid z)P(w\mid z)}{\Sigma_{z'}P(z')P(d\mid z')P(w\mid z')} \quad (11)$$

Maximization Step (M-step):

$$P(w \mid z) = \frac{\Sigma_d n(d,w)P(z\mid d,w)}{\Sigma_{d,w'} n(d,w')P(z\mid d,w')} \quad (12)$$

$$P(d \mid z) = \frac{\sum_w n(d,w)P(z \mid d,w)}{\sum_{d',w} n(d',w)P(z \mid d',w)} \quad (13)$$

$$P(z) = \frac{1}{R}\sum_{d,w} n(d,w)P(z \mid d,w), R \equiv \sum_{d,w} n(d,w) \quad (14)$$

According to Equation (10), It can be seen that PLSA estimates the probability distribution of each document on the latent topics independently. There are $NK+MK$ parameters $\{P(z_j|d),P(w|z_j)\}$ in PLSA model, consequently the number of parameters grow linearly with the number of training documents N. This behavior suggests that PLSI is prone to overfitting (Cai, et al., 2008).

This problem has been addressed in the LDA model. LDA model treats the probability distribution of each document on the latent topics as a *K-parameter* hidden latent variables generated from the same Dirichlet distribution. The generative process of LDA is given below.

- Select a document d, with probability $P(d)$.
- Pick a latent topic $z_{d,m}$:
  - Generate a topic proportion $\theta_d \sim Dir(\alpha)$,
  - Pick a latent topic $z_{d,m}$ with probability $P(z_{d,m} \mid \theta_d)$.
- Generate a word w with probability $P(w_{d,m}|z_{d,m})$.

where, Dir(α) is a K-dimensional Dirichlet distribution. The graphical model for LDA is as shown in Figure 7(b).

Formally, let the topics be $\beta_{1:K}$, where each β_K is a distribution over a set of words ($\beta_k \sim Dir(\eta)$). The topic proportions for the dth document are θ_d, where $\theta_{d,k}$ is the topic proportion for topic k in document d. The topic assignments for the dth document are z_d, where $z_{d,m}$ is the topic assignment for the mth word in document d. And, the observed words for document dth are w_d, where $w_{d,m}$ is the mth word in document d. According to this notation, the LDA generative process corresponds to the following joint distribution of the hidden and observed variables (Blei, et al., 2003).

$$p(\beta_{1:K}, \theta_{1:N}, z_{1:N}, w_{1:N}) = \prod_{j=1}^{K} p(\beta_j) \prod_{d=1}^{D} p(\theta_d) (\prod_{m=1}^{M} p(z_{d,m} \mid \theta_d) p(w_{d,m} \mid \beta_{1:K}, z_{d,m})) \quad (15)$$

The posterior of LDA is

$$p(\beta_{1:K}, \theta_{1:N}, z_{1:N} \mid w_{1:N}) = \frac{p(\beta_{1:K}, \theta_{1:N}, z_{1:N}, w_{1:N})}{p(w_{1:N})} \quad (16)$$

The exact value of this posterior is intractable to compute due to the exponentially large number of possible latent topic structures. Generally, the posterior of LDA can be computed using either sampling based algorithms (e.g. Gibbs sampling, see Heinrich, 2009) or variational algorithms (e.g. coordinate ascent variational inference, see Hoffman, et al., 2010).

Notice that the number of parameters in a *K-topic* LDA is *K+MK*, which does not grow with the number of documents N in the corpus. Consequently, LDA does not have the overfitting issue as PLSA.

Topic Modeling for Web Community Discovery

Recall that, previously in the context of link based community discovery, a Web community is defined as a set of Web pages on a *specific topic*. Now, we turn our discussion to the discovery of Web community based on the Bayesian approach. Let us first re-define some notations and formulate the Web community discovery problem in our new context. This formulation mainly follows the work of Mei et al. (2008).

Problem Formulation

Definition 11 (Document): A *document* d in a text collection D is a bag of words $\{w_1 w_2 ... w_{|d|}\}$, where w_i is a word from a fixed vocabulary. We use $n(d,w)$ to denote the occurrences of word w in d.

Definition 12 (Network): A *network* associated with a text collection D is a graph $G(V, E)$, where V is a set of nodes and E is a set of edges. A node $u \in V$ corresponds to a document $d_u \in D$. An edge $<u,v>$ corresponds to a binary relation between nodes u and v, and $w(u,v)$ is used to denote the weight of $<u,v>$. Note that, an edge can be either directed or undirected.

Definition 13 (Topic): A semantically coherent *topic* in a text collection D is represented by a *topic model* θ, which is a probability distribution of words.

Definition 14 (Topical Web Community Discovery): Given a collection D *and* its associated network structure G, the task of *topical Web community discovery* is to extract k topical com-

munities $\{V_1, \ldots, V_k\}$, where each $V_i = \{v_{i1}, \ldots, v_{il}\}$ has a coherent semantic summary θ_i, which is one of the k major topics in D.

A Survey of Topic Modeling-Based Methods for Web Community Discovery

Traditional topic models, e.g. PLSA (Hofmann, 1999) and LDA (Blei, et al., 2003), analyze content of a given corpus to uncover hidden topics within the corpus. These topics can be naturally interpreted as a community. However, the content analysis alone cannot accurately identify Web communities because the content information usually contains words that are irrelevant to the target topics.

Many link based probabilistic models have been developed for Web community discovery. Cohn et al. (2000) proposed PHITS, a PLSA-like topic model for identification of communities of hubs and authorities in a document network. The PHITS model defines generative processes for both text and hyperlinks. The PHITS model assigns high probability of hyperlinking to a document d with respect to a topic k if the document d is pointed to by several documents that are relevant to the topic k (which are those documents sharing the same word distribution as d). Based on this same assumption, Erosheva et al. (2004) proposed the Link-LDA model that uses LDA as the basic generative building block instead of the PLSA. Recently, Yang et al. (2010) has proposed a probabilistic model for *directed* network community detection, called PPL that captures both incoming links and outgoing links differentially.

Next, let us describe the work that combines link and content analysis into the probabilistic topic modeling framework. The work in this area can be classified into five directions.

The first line of work (e.g. PHITS-PLSA by Cohn & Hofmann, 2001; LDA-Link-Word by Erosheva, et al., 2004; Link-PLSA-LDA by Nallapati, et al., 2008; Dietz, et al., 2007; Gruber, et al., 2008) incorporates the notion of link information into the document generative model. Such methods need expert knowledge in order to translate the semantic of links between documents and embed it into the model, and thus are not generalize to different types of datasets.

The second line of work (relational or supervised topic models) models textual content and link separately by representing the link between documents as a binary random variable conditioned on their content. Prior work in this direction is the Relational Topic Model or RTM by Chang and Blei (2009). Note that, the RTM model does not support weighted graph.

The third line of work regularizes topic models with a discrete regularizer defined based on the link structure of the data set (NetPLSA by Mei, et al., 2008). NetPLSA combines link and text information into a unified framework via the combination of two objective functions, one based on textual data and another one based on the network structure.

The fourth line of work (iTopicModel by Sun, et al., 2009) models the relationship between documents using a multivariate Markov Random Field (MRF). The iTopicModel constructs a two-layer graphical model structure. The top layer is a multivariate MRF that capture the dependency relationship among documents in the network. The bottom layer is a traditional topic model. The assumption underpinning the iTopicModel is the *topical locality* of documents in the network i.e. that the documents in the same neighborhood in the network should be topically similar (Davidson, 2000).

Lastly, the fifth direction, Yang et al. (2009) proposed the PCL model, which is a discriminative model for combining link and content information for community detection.

FUTURE RESEARCH DIRECTIONS

A central problem in the study of semantically coherent implicit Web community that we have discussed so far is how to efficiently extract those communities given a snapshot of a Web dataset. Because the Web is dynamic and constantly changing, it is crucial to understand the dynamic nature of the Web. Therefore, the *temporal* aspect of Web community identification problem is one of important research directions to pursue. Another key research problem is the evaluation of the community discovery algorithms and the *quality* of extracted communities. Mei et al. (2008) compared the performance of NetPLSA with the traditional PLSA using the cut edge weights and ratio cut metrics. However, these metrics only consider the quality based on the link density of the community. Leskovec et al. (2010) proposed an idea to consider the quality of community as a function of its size and conducted large-scale empirical comparison of algorithms for network community detection. However, they only considered non-Bayesian methods.

CONCLUSION

In this chapter, we have explored two complementary approaches to Web community identification: non-Bayesian and Bayesian. The non-Bayesian approaches exploit merely the network structure and extract communities based on some identifying signatures of Web communities. Although they have proved successful, the major limitation of the non-Bayesian based approaches is the lack of semantic interpretation for the extracted communities. The Bayesian approaches on the other hand can be enhanced to combine both textual and link information, and have become popular as algorithmic tools for extracting Web communities from large document network corpora. There is still a lot of opening research issues in this area, and we have identified two interesting future research directions: community dynamics and evaluation of Web community extraction methods.

REFERENCES

Albert, R., Jeong, H., & Barabasi, A. (1999). The diameter of the world wide web. *Nature, 401*(130).

Anderson, R., & Lang, K. (2006). Communities from seed sets. In *Proceedings of the 15th International Conference on World Wide Web (WWW 2006)*. IEEE.

Barabasi, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286.

Bernardo, A. H., Peter, P., James, E. P., & Rajan, M. L. (1998). Strong regularities in world wide web surfing. *Science, 280*(5360), 95–97. doi:10.1126/science.280.5360.95

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Boldi, P., Codenotti, B., Santini, M., & Vigna, S. (2002). Structural properties of the African web. In *Proceedings of the 11th International Conference on World Wide Web (WWW 2002)*. IEEE.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., & Stata, R. ... Wiener, J. (2000). Graph structure in the web. In *Proceedings of the 9th International Conference on World Wide Web (WWW 2000)*. IEEE.

Cai, D., Mei, Q., Han, J., & Zhai, C. (2008). Modeling hidden topics on document manifold. In *Proceedings of ACM 17th Conference on Information and Knowledge Management (CIKM 2008)*. ACM Press.

Chakrabarti, S., Dom, B., Gibson, D., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1998). Experiments in topic distillation. In *Proceedings of the SIGIR Workshop on Hypertext Information Retrieval on the Web*. ACM.

Chang, J., & Blei, D. (2009). Relational topic models for document networks. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*. IEEE.

Chung, F. R. K. (1997). *Spectral graph theory*. New York, NY: AMS Bookstore.

Cohn, D., & Chang, H. (2000). Learning to probabilistically identify authoritative documents. In *Proceedings of the International Conference on Machine Learning (ICML 2000)*. ICML.

Cohn, D., & Hofmann, T. (2001). The missing link - A probabilistic model of document content and hypertext connectivity. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2001)*. NIPS.

Davidson, B. D. (2000). Topical locality in the web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*. ACM Press.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological, 39*, 1–38.

Dhillon, I., Guan, Y., & Kullis, B. (2007). Weighted graph cuts without Eigen vectors: A multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(11), 1944–1957. doi:10.1109/TPAMI.2007.1115

Dhillon, I. S., Mallela, S., & Modha, D. S. (2003). Information-theoretic co-clustering. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003)*. ACM Press.

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the International Conference on Machine Learning (ICML 2007)*. ICML.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 5220–5227. doi:10.1073/pnas.0307760101

Flake, G. W., Lawrence, S., & Gile, C. L. (2000). Efficient identification of web communities. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2000)*. ACM Press.

Flake, G. W., Lawrence, S., Gile, C. L., & Coetzee, F. (2002). Self-organization of the web and identification of communities. *IEEE Computer, 35*(3), 66–71. doi:10.1109/2.989932

Ford, L., & Fulkerson, D. (1956). Maximal flow through a network. *Canadian Journal of Mathematics, 8*(3), 399–404. doi:10.4153/CJM-1956-045-5

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. New York, NY: W. H. Freeman.

Gibson, D., Kleinberg, J., & Raghavan, P. (1998). Inferring web communities from link topology. In *Proceedings of 9th ACM Conference on Hypertext and Hypermedia*. ACM Press.

Gibson, D., Kumar, R., McCurley, K. S., & Tomkins, A. (2006). Dense subgraph extraction. In *Mining Graph Data* (pp. 411–441). New York, NY: Wiley. doi:10.1002/9780470073049.ch16

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. IEEE.

Griffiths, T. L., & Steyvers, M. (2003). Prediction and semantic association. In *Neural Information Processing Systems*. Cambridge, MA: MIT Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 5228–5235. doi:10.1073/pnas.0307752101

Gruber, A., Rosen-Zvi, M., & Weiss, Y. (2008). Latent topic models for hypertext. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI 2008)*. UAI.

Heinrich, G. (2009). *Parameter estimation for text analysis. Technical Report*. Darmstadt, Germany: Fraunhofer IGD.

Hoffman, M., Blei, D., & Bach, F. (2010). Online learning for latent dirichlet allocation. *Neural Information Processing Systems*. Retrieved from http://www.cs.princeton.edu/~blei/papers/HoffmanBleiBach2010b.pdf

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*. IEEE.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal, 42*(1), 177–196. doi:10.1023/A:1007617005950

Hofmann, T., Puzicha, J., & Jordan, M. I. (1999). Advances in Neural Information Processing Systems: *Vol. 11. Unsupervised learning from dyadic data*. Cambridge, MA: MIT Press.

Ino, H., Kudo, M., & Nakamura, A. (2005). Partitioning of web graphs by community topology. In *Proceedings of the 14th International Conference on World Wide Web (WWW 2005)*. IEEE.

Karypis, G., & Kumar, V. (1999). Parallel multilevel k-way partitioning for irregular graphs. *SIAM Review, 41*(2), 278–300. doi:10.1137/S0036144598334138

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms*. ACM Press.

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the web for emerging cyber-communities. In *Proceedings of the 8th International Conference on World Wide Web*. IEEE.

Langville, A. N., & Meyer, C. D. (2006). *Google's PageRank and beyond: The science of search engine rankings*. Princeton, NJ: Princeton University Press.

Leskovec, J., Lang, K. J., & Mahoney, M. W. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*. IEEE.

Li, L., Otsuka, S., & Kitsuregawa, M. (2010). Finding related search engine queries by web community based query enrichment. *World Wide Web (Bussum), 13*(1-2), 121–142. doi:10.1007/s11280-009-0077-1

Liu, B. (2007). *Web data mining: Exploring hyperlinks, contents, and usage data*. Berlin, Germany: Springer.

Mei, Q., Deng, C., Zhang, D., & Zhai, C. (2008). Topic modeling with network regularization. In *Proceedings of the 17th International Conference on World Wide Web (WWW 2008)*. IEEE.

Nallapati, R. M., Ahmed, A., Xing, E. P., & Cohen, W. W. (2008). Joint latent topic models for text and citations. In *Proceedings of The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*. ACM Press.

Otsuka, S., Toyoda, M., Hirai, J., & Kitsuregawa, M. (2004). Extracting user behavior by web communities technology on global web logs. In *Proceedings of the 15th International Conference on Database and Expert Systems Applications (DEXA 2004)*. DEXA.

Pierrakos, D., Paliouras, G., Papatheodorou, C., Karkaletsis, V., & Dikaiakos, M. D. (2003). Web community directories: A new approach to web personalization. In *Proceedings of the 1st European Web Mining Forum (EWMF 2003)*. EWMF.

Somboonviwat, K. (2008). *Research on language specific crawling and building of Thai web archive.* (Unpublished Doctoral Dissertation). University of Tokyo. Tokyo, Japan.

Sun, Y., Han, J., Gao, J., & Yu, Y. (2009). iTopicModel: Information network-integrated topic modeling. In *Proceedings of 2009 International Conference on Data Mining (ICDM 2009)*. ICDM.

Tamura, T., Somboonviwat, K., & Kitsuregawa, M. (2007). A method for language-specific web crawling and its evaluation. *Systems and Computers in Japan, 38*(2), 10–20. doi:10.1002/scj.20693

Toyoda, M., & Kitsuregawa, M. (2001). Creating a web community chart for navigating related communities. In *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia (HT 2001)*. ACM Press.

Toyoda, M., & Kitsuregawa, M. (2003). Extracting evolution of web communities from a series of web archives. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia (HT 2003)*. ACM Press.

Yang, T., Chi, Y., Zhu, S., & Jin, R. (2010). Directed network community detection: A popularity and productivity link model. In *Proceedings of the 2010 SIAM International Conference on Data Mining (SDM 2010)*. SIAM.

Yang, T., Jin, R., Chi, Y., & Zhu, S. (2009). Combining link and content for community detection: A discriminative approach. In *Proceedings of The 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009)*. ACM Press.

Yu, S., Moor, B. D., & Moreau, Y. (2009). Clustering by heterogeneous data fusion: Framework and applications. In *Proceedings of the NIPS Workshop*. NIPS.

Zhang, Y., Xu Yu, J., & Hou, J. (2006). *Web communities: Analysis and construction*. Berlin, Germany: Springer.

Zhu, S., Yu, K., Chi, Y., & Gong, Y. (2007). Combining content and link for classification using matrix factorization. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*. ACM Press.

ข้อมูลประวัติคณะผู้วิจัย

ประวัติส่วนตัว

ชื่อ-สกุล กุลวดี สมบูรณ์วิวัฒน์

เพศ ☐ ชาย ☑ หญิง วันเดือนปีเกิด 17 มิถุนายน 2522 อายุ 33 ปี

สถานภาพ ☑ โสด ☐ สมรส

ตำแหน่งปัจจุบัน อาจารย์

ประวัติการศึกษา

ชื่อย่อปริญญา	สาขา	สถาบันที่จบ	ปีที่จบ
วศบ. (เกียรตินิยมอันดับ 1)	วิศวกรรมคอมพิวเตอร์	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง	2543
M.Sc.	Information and Communication Engineering	The University of Tokyo	2548
Ph.D.	Information and Communication Engineering	The University of Tokyo	2551

สาขาวิจัยที่มีความชำนาญพิเศษ (แตกต่างจากวุฒิการศึกษา) Database, Web mining, Machine learning, Bioinformatics

รางวัลด้านวิชาการ/ด้านวิจัย/งานสร้างสรรค์ (ด้านศิลปะ หรืออื่นๆ) ที่ได้รับ

ปี พ.ศ.	ชื่อรางวัล	สถาบันที่ให้
2555	Hitachi Research Fellowship 2012	The Hitachi Scholarship Foundation
2550	Best Presentation Award	DBWeb2007, Tokyo, Japan

ทุนการศึกษาและทุนวิจัยที่เคยได้รับ

ปี พ.ศ.	ทุนการศึกษาและทุนวิจัย	สถาบันที่ให้
2548-2551	Tonen International Scholarship	Tonen International Scholarship Foundation
2545-2548	Panasonic Scholarship	Panasonic Scholarship Inc.
2543	SCG Talent Scholarship	The Siam Cement Group

ผลงานวิจัย/งานสร้างสรรค์

ผลงานวิจัย/งานสร้างสรรค์ที่ตีพิมพ์เผยแพร่/การเสนอผลงานวิชาการ

PhD Thesis

[1] Kulwadee Somboonviwat, 2008. Research on Language Specific Crawling and Building of Thai Web Archive. Thesis (PhD), The University of Tokyo.

Journal/Edited Book Chapter

[2] Kulwadee Somboonviwat. *Web Community Discovery and Topic Modeling.* Social Media Mining and Social Network Analysis: Emerging Research, IGI-Global, 2013.

[3] Takayuki Tamura, Kulwadee Somboonviwat, and Masaru Kitsuregawa. A method for language-specific Web crawling and its evaluation. Systems and Computers in Japan, 38(2):10-20, 2007.

Peer-reviewed International Conference

[4] Kulwadee Somboonviwat. *Web Community Analysis and its Application to Language Specific Crawling.* in Proc. of the 6th International Conference on Ubiquitious Information Management and Communication (ICUIMC 2012), 2012.

[5] Kulwadee Somboonviwat, Shinji Suzuki, and Masaru Kitsuregawa. Connectivity of the Thai Web Graph.In Proc. of 10th Asia-Pacific Web Conference, (APWeb 2008), pp. 613-624, 2008.

[6] Kulwadee Somboonviwat, Shinji Suzuki, and Masaru Kitsuregawa. Structure of the Thai Web Graph. In Proc. of the 2008 IEEE International Symposium on Mining And Web (IEEE MAW-08), 2008.

[7] Kulwadee Somboonviwat, Shinji Suzuki, Masashi Toyoda, and Masaru Kitsuregawa. Characterization of the Thai Hostgraph. In Proc. of the Second International Conference on Ubiquitous Information Management and Communication (ICUIMC 2008), pp. 376-381, 2008.

[8] Lin Li, Zhenglu Yang, Kulwadee Somboonviwat, and Masaru Kitsuregawa: User-assisted similarity estimation for searching related web pages. In Proc. of Hypertext 2007, pp. 11-20, 2007.

International Workshop

[9] Kulwadee Somboonviwat, Shinji Suzuki, and Masaru Kitsuregawa. A Preliminary Study on the Extraction of Socio-topical Web Keywords. In Proc. of the International Workshop on Scalable Web Information Integration and Service (SWIIS2007), pp.74-82, 2007.

[10] Kulwadee Somboonviwat, Takayuki Tamura, and Masaru Kitsuregawa. Finding Thai Web Pages in Foreign Web Spaces. In Proc. of the Second International Special Workshop on Databases For Next Generation Researchers In Memoriam Prof. Yahiko Kambayashi (SWOD2006), pp.130-133, 2006.

[11] Kulwadee Somboonviwat, Takayuki Tamura, and Masaru Kitsuregawa. Simulation Study of Language Specific Web Crawling. In Proc. of the International Special Workshop on Databases For Next Generation Researchers In Memoriam Prof. Yahiko Kambayashi (SWOD2005), pp.142-145, 2005.

Regional Conference

[12] Kulwadee Somboonviwat, Kunlaya Somboonwiwat, Sureerat Tang, Anchalee Tassanakajon. Architecture and Implementation of a Large-scale BATCH BLAST Searches on a Cluster of Workstations. In Proc. of the 3rd Regional Conference on ICT Applications for Industries and Small Companies in ASEAN countries (RCICT 2011), 2011.

Domestic Workshop

[13] Kulwadee Somboonviwat, and Masaru Kitsuregawa. Inferring Link Behavior from the Connectivity Distributions of Web Pages. In Proc. of 19th Data Engineering Workshop (DEWS2008), A1-1, 2008.

[14] Kulwadee Somboonviwat, Shinji Suzuki, Masashi Toyoda, and Masaru Kitsuregawa. Thematic and Temporal Analysis of Thai Web Communities. The 70th National Convention of IPSJ, 6J-1, 2008.

[15] Kulwadee Somboonviwat, Takayuki Tamura, Masaru Kitsuregawa. Finding Thai Web Pages in Foreign Web Spaces. Proceedings of Data Engineering Workshop (DEWS2006), 3A-i6, 2006.

[16] Kulwadee Somboonviwat, Takayuki Tamura, Masaru Kitsuregawa. Simulation Study of Language Specific Web Crawling, Proceedings of Data Engineering Workshop (DEWS2005), 4B-o1, 2005.

ผลงานสิทธิบัตร/สิ่งประดิษฐ์/งานสร้างสรรค์ (ศิลปะ หรือ อื่นๆ)

-

อื่นๆ

-