วิทยานิพนธ์

การปรับลดอิทธิพลบูซฟาร์มในการคำนวณเพจแร็งค์

UN-BIASING BOOST FARM IN PAGERANK COMPUTATION

นายกำธร พันธุมะผล

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

พ.ศ. 2550

วิทยานิพนธ์

เรื่อง

การปรับลดอิทธิพลบูซฟาร์มในการคำนวณเพจแร็งค์

Un-Biasing Boost Farm in PageRank Computation

โดย

นายกำธร พันธุมะผล

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
พ.ศ. 2550

กำธร พันธุมะผล 2550: การปรับลดอิทธิพลบูซฟาร์มในการคำนวณเพจแร็งค์
ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: ผู้ช่วยศาสตราจารย์อานนท์ รุ่งสว่าง,
Ph.D. 97 หน้า

ในปัจจุบันการวิเคราะห์ลิงค์เป็นองค์ประกอบสำคัญที่ใช้ในการเรียงลำดับผลลัพธ์การค้นคืน และเชื่อว่าถูกใช้โดยระบบสืบค้นข้อมูลชื่อดังหลายแห่ง อาทิเช่น ระบบสืบค้นข้อมูลกูลเกิ้ล (Google 2007) และยาฮู(Yahoo 2007) ซึ่งผลลัพธ์จากระบบสืบค้นข้อมูลนั้นเสมือนเป็นจุดเริ่มต้นอันนำไปสู่เว็บเพจที่นับวันจะมีความสำคัญทางธุรกิจมากขึ้น จึงทำให้มีความพยายามในการพัฒนาการสแปมโครงสร้างลิงค์ของเว็บ เพื่อให้เว็บไซต์ของตนนั้นถูกจัดลำดับโดยระบบสืบค้นข้อมูลในลำดับที่สูงที่สุด ซึ่งการกระทำในลักษณะนี้ส่งผลให้ผลค้นคืนจากระบบสืบค้นข้อมูลไม่น่าเชื่อถือ และเราจะเรียกการกระทำในลักษณะนี้ว่า “การสแปมระบบสืบค้นข้อมูล” วิธีการหนึ่งในการสแปมระบบสืบค้นข้อมูลคือการสร้างโครงสร้างลิงค์ที่มีการเชื่อมต่ออย่างหนาแน่นที่เรียกว่า “บูซฟาร์ม” นั้นเป็นเทคนิคหนึ่งที่สามารถจะส่งผลกระทบต่อการจัดลำดับความสำคัญที่ใช้โครงสร้างลิงค์เป็นหลักได้ จากปัญหาที่กล่าวมาในวิทยานิพนธ์นี้จึงนำเสนอวิธีการปรับลดอิทธิพลของบูซฟาร์มในการคำนวณเพจแร็งค์ ซึ่งเป็นการจัดลำดับเว็บเพจโดยการวิเคราะห์ลิงค์โดยการสร้างโครงสร้างลิงค์เสมือนเพื่อแก้ไขโครงสร้างลิงค์ของบูซฟาร์มพร้อมทั้งนำเสนอผลการทดลองที่น่าสนใจในเบื้องต้น

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Komthorn Puntumapon 2007: Un-Biasing Boost Farm in PageRank Computation. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering.
Thesis Advisor: Assistant Professor Arnon Rungsawang,
Ph.D. 97 pages.

Link analysis is one of the important components in the current search engine's ranking system e.g. Google and Yahoo. Since the search result leads to the web pages that will later have implicit effect in business process, there exist many attempts to spam the link structure of the web for boosting up web's ranking. This behavior degrades the search engine's results and we called this behavior "search engine spam". Building an artificial condensed link structure called "boost farm" is a technique that can optimize the link based ranking system. In this paper, we proposed an approach to un-bias the effect of those boost farms in the PageRank computation by creating virtual links for correcting boost farm's link structure, as well as the promising preliminary results.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

ข้าพเจ้าขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง ประธานกรรมการที่ปรึกษา ที่ได้ช่วยเหลือในการวางแผนงานวิจัยในวิทยานิพนธ์ฉบับนี้ ตลอดจนการให้คำปรึกษา พร้อมทั้งให้แนวทางและความรู้เกี่ยวกับทฤษฎีต่างๆ มากมายในการทำวิจัย รวมถึงข้อเสนอแนะที่เป็นประโยชน์ต่อการทำวิทยานิพนธ์ฉบับนี้ รวมถึงการตรวจแก้ไขข้อบกพร่องต่างๆ และขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.อนันต์ ผลเพิ่ม ที่กรุณาให้คำปรึกษาแนะนำและได้ให้ข้อเสนอแนะดีๆ ในการทำวิทยานิพนธ์ให้สำเร็จลุล่วงไปด้วยดี

ขอขอบคุณอาจารย์ อุดมพร ตุงคะศิริ ที่ให้คำปรึกษาที่ดีมาโดยตลอด ไม่ว่าจะเป็นทิศทางของงานวิจัย ความรู้ต่างๆ มากมาย และขอขอบคุณพี่บัณฑิต มนัสเกษมศักดิ์ ที่คอยช่วยให้คำปรึกษาไม่ว่าจะเป็นเรื่องงานและเรื่องอื่นๆมากมาย และสุดท้ายขอขอบคุณสมาชิกห้องปฏิบัติการ MIKE ทุกท่านที่คอยให้คำแนะนำและกำลังใจที่ดีเสมอมา

ขอขอบคุณพี่ชัชชฎา รัตนกิจภิญโญ เจ้าหน้าที่ธุรการโครงการปริญญาโทที่ช่วยเหลือในการประสานงานและงานด้านเอกสารต่างๆ ให้งานเป็นไปอย่างสะดวกลุล่วงไปด้วยดี รวมถึงขอขอบคุณเจ้าหน้าที่ธุรการภาควิชาวิศวกรรมคอมพิวเตอร์มหาวิทยาลัยเกษตรศาสตร์ทุกท่าน

คุณงามความดี หรือประโยชน์อันใดที่เกิดจากวิทยานิพนธ์ฉบับนี้ ขออุทิศให้แก่ บิดา มารดา บุพการี และผู้มีพระคุณทุกท่าน

กำธร พันธุมะผล

ธันวาคม 2549

สารบัญ

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ (ต่อ)

คำอธิบายสัญลักษณ์และคำย่อ

ACB	=	Average Change rate of probability of Boost farm
APR	=	Average PageRank of good supporter
AGL	=	Average number of Good inLinks
BFM	=	Boost Farm Matrix
non_BFM	=	non-Boost Farm Matrix
PR	=	PageRank
URPC	=	Using Rank Propagation and Probabilistic Counting for Link Based Spam Detection
VM	=	Virtual link Matrix

การปรับลดอิทธิพลบูซฟาร์มในการคำนวณเพจแร็งค์

Un-Biasing Boost Farm in PageRank Computation

คำนำ

ปัจจุบันระบบสืบค้นข้อมูล (search engine) เป็นเครื่องมืออันทรงอานุภาพในการนำพาผู้ใช้ไปยังเว็บเพจต่างๆ โดยปกติแล้วผลลัพธ์ที่ได้จากการสืบค้นทั้งหมดจากคำถามใดคำถามหนึ่งจะมีเพียง 10 เว็บเพจแรกเท่านั้นที่ผู้ใช้งานระบบสืบค้นข้อมูลจะสนใจเปิดอ่านรายละเอียดภายในเว็บเพจนั้นๆ (Henzinger *et al.,* 2002) ซึ่งการเปิดเข้าใช้งานเว็บเพจทางธุรกิจมีผลกระทบต่อการสร้างรายได้ ดังนั้นผู้พัฒนาเว็บเพจทางธุรกิจจึงมีความต้องการให้เว็บเพจของตนเองถูกจัดลำดับใน 10 ลำดับแรก หรือถูกจัดลำดับ (ranking) ในตำแหน่งที่สูงที่สุดเท่าที่จะทำได้

เว็บเพจที่มีคุณภาพสูงมักจะถูกจัดลำดับโดยระบบสืบค้นข้อมูลในลำดับต้นๆ ซึ่งการที่ระบบสืบค้นข้อมูลจัดลำดับเว็บเพจนั้นมักจะพิจารณาจากโครงสร้างของลิงค์ (link structure) ที่เชื่อมต่อกันระหว่างเว็บเพจเป็นหลัก (Markus, 2003) จึงทำให้มีผู้พัฒนาเว็บเพจ ทางธุรกิจบางกลุ่มพยายามสร้างโครงสร้างลิงค์เพื่อให้เว็บเพจถูกคำนวณจากระบบสืบค้นข้อมูลให้อยู่ในลำดับต้นๆได้ ซึ่งการกระทำในลักษณะนี้จะเข้าข่ายที่เรียกว่า “การสแปมระบบสืบค้นข้อมูล” (search engine spam) ในปัจจุบันผู้พัฒนาระบบสืบค้นข้อมูลได้วิจัย และพัฒนาวิธีการต่างๆในการค้นหากลุ่มเว็บเพจที่อยู่ในโครงสร้างที่ถูกสร้างขึ้นเหล่านั้น (Wu and Brian, 2005a; Becchetti *et al.,* 2006) ในการค้นหากลุ่มเว็บเพจเหล่านั้น เรามักจะได้กลุ่มของเว็บเพจอื่นๆที่มีโครงสร้างลิงค์ที่คล้ายคลึงกับการทำสแปมระบบสืบค้นข้อมูลติดมาด้วย เมื่อผู้ดูแลระบบได้ตรวจพบเว็บเพจเหล่านั้นก็จะทำการตัดเว็บเพจดังกล่าวออกจากการจัดลำดับผลลัพธ์ ทั้งๆที่บางเว็บเพจอาจจะยังเป็นคำตอบที่ดีสำหรับคำถามบางคำถามก็เป็นได้ หรืออาจกล่าวได้ว่าระบบสืบค้นข้อมูลระบบนั้นๆอาจได้ตัดบางส่วนของคำตอบออกจากการตอบคำถาม ทำให้คำตอบที่ได้เป็นคำตอบที่ไม่สมบูรณ์

วิธีการค้นหาการทำสแปมระบบสืบค้นข้อมูล ได้ถูกศึกษาโดยนักวิจัยหลายๆท่าน โดยมีการจัดแบ่งรูปแบบออกเป็น 2 ลักษณะใหญ่ๆคือ การสแปมเนื้อหาเอกสารของเว็บเพจ และการสแปมโครงสร้างลิงค์ของเว็บเพจ (Gyongyi and Garcia-Molina 2005a) สืบเนื่องจากความสำเร็จของวิธีการจัดลำดับเว็บเพจโดยพิจารณาจากความสัมพันธ์ของลิงค์ เช่น อัลกอลิทึมฮิตส์ (Kleinberg,

1999) และเพจแร็งค์ (Page *et al*., 1998; Markus, 2003) ทำให้วิธีการสแปมในปัจจุบันมุ่งเน้นไปยังวิธีการสแปมโครงสร้างลิงค์ของเว็บเพจเป็นหลัก เนื่องจากมีนักวิจัยหลายๆท่านได้นำเสนอวิธีในการค้นหากลุ่มของเว็บเพจเหล่านั้นโดยอัตโนมัติแล้ว(Wu and Brian, 2005a; Benczur and Csalogany, 2005; Becchetti *et al*., 2006; Markus, 2003) ดังนั้นในงานวิจัยฉบับนี้ เราจึงให้ความสำคัญเฉพาะวิธีลดผลกระทบของการสแปมโครงสร้างลิงค์ของเว็บเพจเป็นหลัก

ในงานวิจัยชิ้นนี้เราจะเรียกกลุ่มของเว็บเพจที่ถูกสร้างขึ้นให้มีโครงสร้างเชื่อมต่อกันหนาแน่นเพื่อยังประโยชน์ให้เว็บเพจใดเว็บเพจหนึ่งถูกคำนวณโดยระบบสืบค้นข้อมูลแล้วให้อยู่ในลำดับต้นๆเหล่านั้นว่า “บูซฟาร์ม” (boost farm) ส่วนวิธีการที่เราใช้ปรับลดอิทธิพลของบูซฟาร์มนั้น ในขั้นตอนแรกเราจะทำการคำนวณค่าอัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์ม หลังจากนั้นเราจะเสนอวิธีการสร้างลิงค์เสมือนให้กับกลุ่มของเว็บเพจในบูซฟาร์มนั้นและใช้อัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์มมาเพื่อใช้ปรับลดผลกระทบของบูซฟาร์ม ขั้นตอนต่อมา เราจะเสนอการสร้างเมตริกซ์ความสัมพันธ์ที่ใช้แสดงโครงร้างลิงค์ใหม่เพื่อใช้ในการคำนวณเพจแร็งค์ โดยที่ยังสามารถใช้วิธีการคำนวณเพจแร็งค์แบบดั้งเดิมในการคำนวณค่าเพจแร็งค์ในเมตริกซ์ความสัมพันธ์ใหม่ที่สร้างขึ้นได้

จากการทดลองในเบื้องต้นกับวิธีที่นำเสนอในงานวิจัยนี้ โดยใช้โครงสร้างเว็บกราฟขนาดประมาณ 250000 เว็บเพจที่ถูกสแปมโดยผู้วิจัยเอง พบว่าวิธีการที่นำเสนอสามารถปรับลดค่าคะแนนเพจแร็งค์เว็บเพจในกลุ่มของบูซฟาร์มที่มีลิงค์ชี้จากเว็บเพจคุณภาพต่ำจำนวนมากให้สามารถถูกจัดลำดับในลำดับที่ต่ำได้ ต่างจากเว็บเพจในกลุ่มของบูซฟาร์มที่มีลิงค์ชี้จากเว็บเพจที่มีคุณภาพจำนวนมากถึงแม้จะถูกปรับลดด้วยวิธีการที่นำเสนอก็ยังสามารถถูกจัดในลำดับที่สูงได้

วัตถุประสงค์

1. พัฒนาเทคนิคการสร้างโครงสร้างลิงค์เสมือนเพื่อปรับลดอิทธิพลของบูซฟาร์มโดยเสนอวิธีการใหม่ในการสร้างลิงค์เสมือนที่ใช้กับกลุ่มของบูซฟาร์มนั้นๆ

2. นำเสนอวิธีการคำนวณค่าอัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์มสำหรับกลุ่มของบูซฟาร์มใดๆเพื่อปรับลดอิทธิพลจากโครงสร้างลิงค์ของบูซฟาร์มเหล่านั้น

การตรวจเอกสาร

ความรู้พื้นฐานและงานวิจัยที่เกี่ยวข้อง

ในหัวข้อนี้จะกล่าวถึงเทคนิคการคำนวณเพจแร็งค์และฮิตส์ ซึ่งเป็นเทคนิคการจัดลำดับเว็บเพจโดยการวิเคราะห์ลิงค์ และงานวิจัยอื่นๆที่เกี่ยวข้อง ยกตัวอย่างเช่น วิธีการจัดประเภทการสแปมระบบสืบค้นข้อมูล วิธีการค้นหาเว็บเพจที่ถูกสแปมระบบสืบค้นข้อมูล

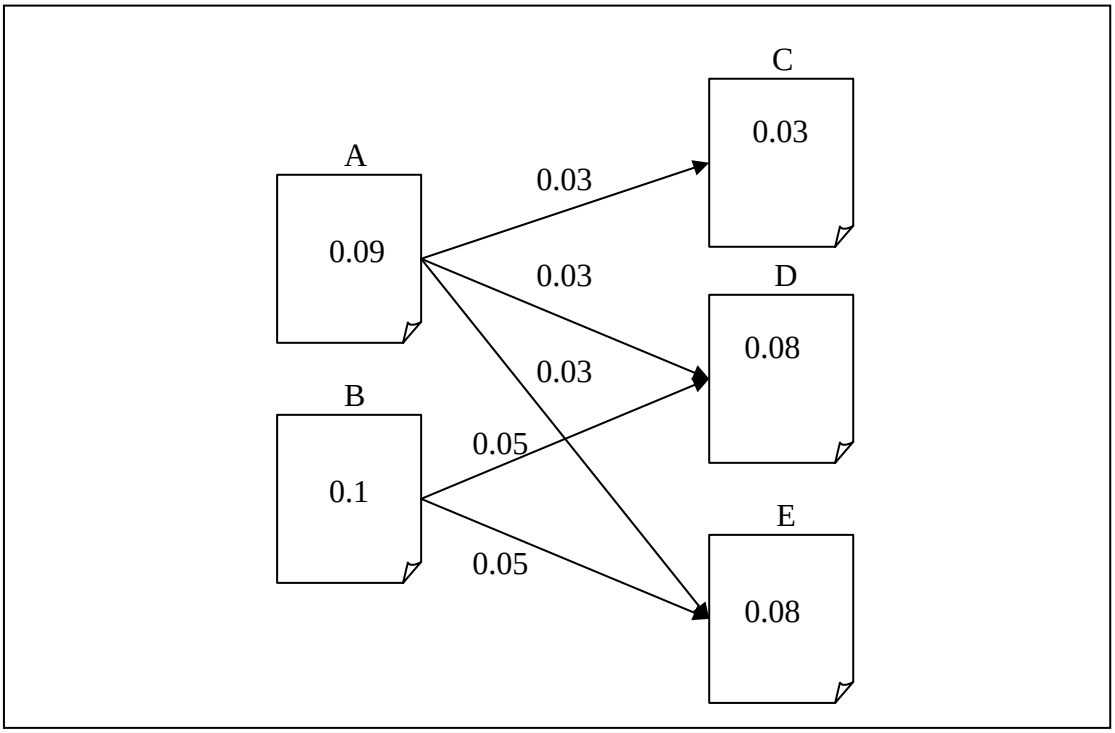
เพจแร็งค์ (PageRank)

เพจแร็งค์เป็นเทคนิคการจัดลำดับเว็บเพจจากผลค้นคืนของระบบสืบค้นข้อมูลโดยการวิเคราะห์โครงสร้างลิงค์ *ใน* ค.ศ. 1998 Page และ Brin ได้สังเกตว่าเว็บเพจในอินเทอร์เน็ตนั้นมีความหลากหลายทางด้านจำนวนลิงค์ที่ชี้มายังเว็บเพจนั้นๆ และจำนวนลิงค์ที่ชี้มายังเว็บเพจเหล่านั้นยังสามารถแสดงถึงความสำคัญของเว็บเพจเหล่านั้นยกตัวอย่างเช่น ในปี ค.ศ. 1998 เว็บเพจ ของ Netscape มีลิงค์ชี้เข้าเป็นจำนวนถึง 62804 ลิงค์ ซึ่งนักวิจัยกล่าวว่าเป็นข้อมูลจากฐานข้อมูล ณ เวลานั้น เมื่อนำมาเปรียบเทียบกับเว็บเพจอื่นๆที่มีจำนวนลิงค์ที่ชี้มายังเว็บเพจนั้นเป็นจำนวนน้อยกว่าจะพบว่าเว็บเพจที่มีจำนวนลิงค์ที่ชี้เข้ามายังเว็บเพจนั้นเป็นจำนวนมากจะเป็นเว็บเพจที่มีความสำคัญเช่นเดียวกันกับการนับจำนวนการอ้างอิงของบทความเพื่อใช้คาดว่าบทความใดควรจะได้รับรางวัลโนเบล ดังนั้นถ้าเว็บเพจต้นทางมีลิงค์ชี้ไปยังเว็บเพจปลายทางหมายความว่าผู้พัฒนาเว็บเพจต้นทางนั้นได้ส่งมอบความสำคัญให้กับเว็บเพจปลายทางนั้นๆ

ถ้ากำหนดให้ $\omega(u)$ แทนจำนวนลิงค์ที่ชี้ออกจากเว็บเพจ u ใดๆ B_v แทนเซตของเว็บเพจที่ชี้ไปยังเว็บเพจ v และ $\mathrm{Rank}_i(u)$ แสดงถึงค่าความสำคัญของเว็บเพจ u ใดๆในการคำนวณเพจแร็งค์ในรอบที่ i ดังนั้นเมื่อเว็บเพจ u มีลิงค์ชี้ไปยังเว็บเพจ v จะหมายความถึงเว็บเพจ u ส่งมอบความสำคัญให้กับเว็บเพจ v ซึ่งสามารถเขียนแทนด้วยสมการดังต่อไปนี้

$$\mathrm{Rank}_{i+1}(v) = \sum_{u \in B_v} \mathrm{Rank}_i(u) / \omega(u) \qquad (1)$$

จากสมการที่ (1) เราจะเห็นได้ว่า Page และ Brin ให้ค่าความสำคัญของเว็บเพจที่ส่งผ่านไปตามลิงค์ที่ชี้ออกไปยังเว็บเพจอื่นๆด้วยค่าเท่าๆกัน ยกตัวอย่างจากภาพที่ 1 แสดงถึงการส่งมอบค่าความสำคัญจากเว็บเพจต้นทางไปยังเว็บเพจปลายทาง



ภาพที่ 1 ตัวอย่างการส่งมอบค่าคะแนนเพจแร็งค์

จากภาพที่ 1 จะเห็นได้ว่าเว็บเพจ A มีค่าเพจแร็งค์เท่ากับ 0.09 จะส่งมอบค่าคะแนนเพจแร็งค์เท่าๆกันไปยังเว็บเพจปลายทาง C, D และ E ด้วยค่าเท่ากับ 0.03 ซึ่งได้มาจากค่าคะแนนของเว็บเพจต้นทางหารด้วยจำนวนลิงค์ทั้งหมดที่ชี้ออกจากเว็บเพจนั้นๆ และเช่นเดียวกันเว็บเพจ B มีค่าเพจแร็งค์เท่ากับ 0.1 จะส่งมอบค่าคะแนนไปยังไปยังเว็บเพจปลายทาง D และ E ด้วยค่าเท่ากับ 0.05 ทำให้เว็บเพจ C, D และ E มีค่าคะแนนเพจแร็งค์ 0.03, 0.08 และ 0.08 ตามลำดับ

จากวิธีการที่กล่าวมานั้นเป็นการคำนวณค่าเพจแร็งค์สำหรับเว็บเพจใดๆ ซึ่งนำไปสู่การคำนวณค่าความสำคัญของเว็บเพจทั้งหมดได้โดยเราจะพิจารณาเว็บเพจบนอินเทอร์เน็ตเป็นเว็บกราฟที่มีทิศทาง $G = (\mathcal{V}, \mathcal{E})$ โดยกำหนดให้ $\mathcal{V}$ คือเซตของเว็บเพจในเว็บกราฟ G และ $\mathcal{E}$ คือเซตของลิงค์ที่เชื่อมต่อระหว่างเว็บเพจ เราจะสามารถเขียนเมตริกซ์ความสัมพันธ์ M_e เป็นตัวแทนโครงสร้างความสัมพันธ์ของเว็บเพจต่างๆเหล่านั้นได้ดังสมการต่อไปนี้

$$M_e(\mathrm{p},\ \mathrm{q}) = \begin{cases} 1/\omega(q) & if\ \ (q, p) \in \varepsilon \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

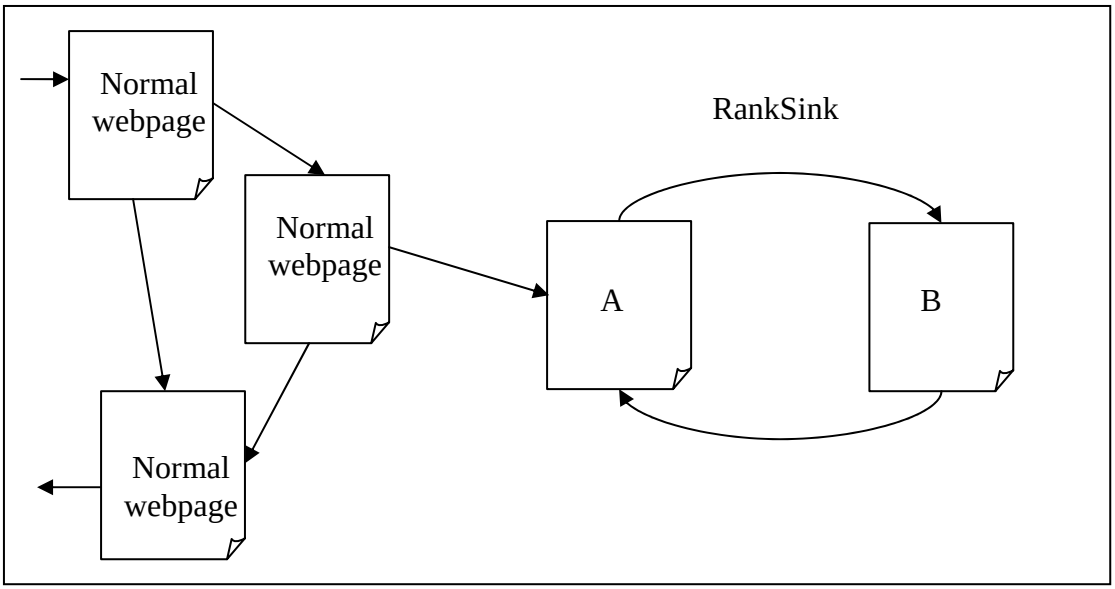
และถ้าเขียนค่าคะแนนของเว็บเพจทั้งหมดในรอบการคำนวณ i ด้วยเวกเตอร์ $\vec{R}_i$ ขนาด $N \times 1$ เมื่อ $N = |V|$ โดยกำหนดให้ค่าคะแนนเพจแร็งค์เริ่มต้นสำหรับทุกเว็บเพจมีค่าเท่ากับ $1/N$ เราจะสามารถเขียนเวกเตอร์ $\vec{R}_0$ เริ่มต้นมีค่าเท่ากับ $[1/N]_{N\times 1}$ จากสมการที่ (1) และ (2) เราสามารถเขียนสมการการคำนวณค่าความสำคัญของเว็บเพจทั้งหมดภายในเว็บกราฟได้ใหม่ดังสมการต่อไปนี้

$$\vec{R}_{i+1} = M_e \vec{R}_i \qquad (3)$$

ในทางปฏิบัติแล้ว เราจะคำนวณสมการที่ (3) แบบวนซ้ำ (iterative computing) จนกว่าเวกเตอร์ $\vec{R}_{i+1}$ ลู่เข้าสู่ค่าซึ่งเป็นคำตอบ (convergence) โดยในแต่ละรอบของการคำนวณนั้น เราสามารถตรวจสอบการลู่เข้าของ $\vec{R}_{i+1}$ ได้ตามสมการที่ (4) กล่าวคือ ถ้าอัตราส่วนความคาดเคลื่อนของเวกเตอร์ (ผลต่างระหว่าง $\vec{R}_{i+1}$ และ $\vec{R}_i$) ต่อขนาดเวกเตอร์ $\vec{R}_i$ มีค่าน้อยกว่าค่าที่ยอมรับได้ δ (threshold) ค่าหนึ่ง เราก็จะพิจารณาว่าเวกเตอร์ $\vec{R}_{i+1}$ ลู่เข้าสู่ค่าซึ่งเป็นคำตอบและหยุดการคำนวณ

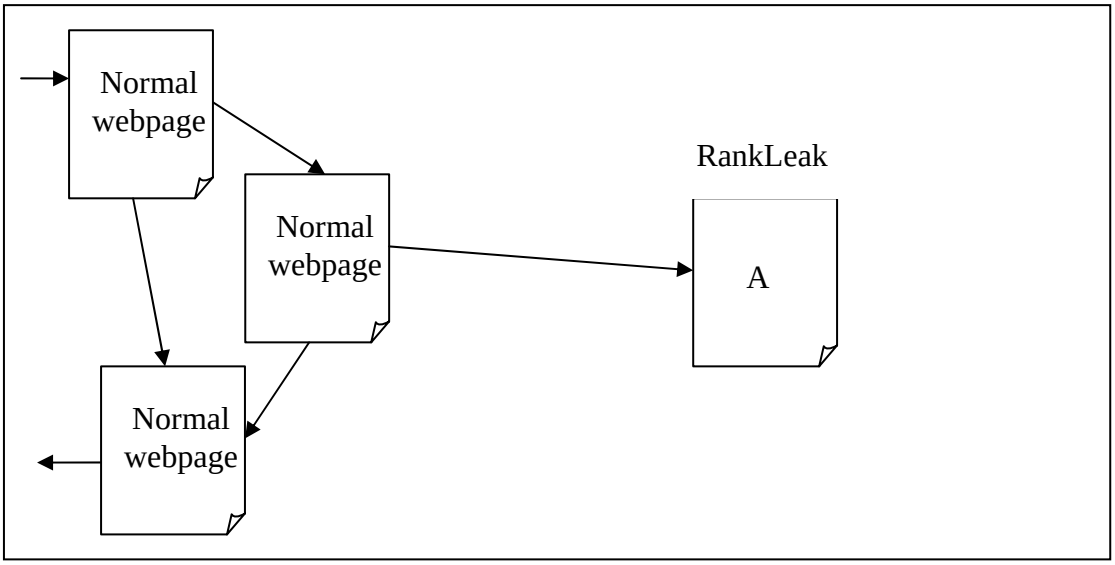
$$\frac{\left\|\vec{\mathrm{R}}_{\mathrm{i}+1} - \vec{\mathrm{R}}_{\mathrm{i}}\right\|_1}{\left\|\vec{\mathrm{R}}_{\mathrm{i}}\right\|_1} \le \delta \qquad (4)$$

แต่อย่างไรก็ตามยังพบปัญหาในการคำนวณคือเมื่อพิจารณากลุ่มของเว็บเพจใดๆที่ลิงค์ชี้เข้าหากันแต่ไม่มีลิงค์ชี้ไปยังเว็บเพจอื่นๆ ในการคำนวณแบบวนซ้ำในสมการที่ (3) นั้นเว็บเพจในกลุ่มนั้นจะเก็บสะสมค่าคะแนนเพจแร็งค์ และไม่กระจายค่าคะแนนกลับสู่ระบบ ในกรณีนี้เราจะเรียกปัญหานี้ว่า แร็งค์ซิงค์ (ranksink) ดังภาพที่ 2



ภาพที่ 2 แร็งค์ซิงค์

และในกรณีที่เว็บเพจใดๆนั้นไม่มีลิงค์ชี้ออกเว็บเพจนั้น เว็บเพจดังกล่าวก็จะได้รับค่าคะแนนเพจแร็งค์จากเว็บเพจที่มีลิงค์ชี้มาแต่ไม่มีการส่งค่าให้กับเว็บเพจอื่นๆ ทำให้ค่าคะแนนเพจแร็งค์นั้นหายออกจากระบบ เราจะเรียกปัญหานี้ว่าแร็งค์ลีค (rankleak) ดังภาพที่ 3



ภาพที่ 3 แร็งค์ลีค

จากปัญหาที่กล่าวมายังผลให้เมตริกซ์ความสัมพันธ์ M_e ไม่สอดคล้องกลับคุณสมบัติของมาคอร์ฟ (Markov's properties) (Kleinrock, 1975) ทำให้คำนวณสมการที่ (3) แบบวนซ้ำเวกเตอร์ $\vec{R}_{i+1}$ อาจจะไม่ลู่เข้าสู่ค่าซึ่งเป็นคำตอบดังนั้น Page และ Brin จึงแก้ไขโดยสร้างเทเลพอตเวกเตอร์ (teleport vector, $\vec{E}$) ซึ่งมีค่าเท่ากับ $[1/N]_{N\times 1}$ ที่หมายความถึงเมื่อเดินทางตามลิงค์ไปถึงเว็บเพจใดๆก็จะมีความน่าจะเป็นค่าเท่ากับ $1/N$ ที่จะเลือกกระโดดไปยังทุกๆเว็บเพจในเว็บกราฟ และเมตริกซ์ความสัมพันธ์ M ที่เป็นตัวแทนโครงสร้างความสัมพันธ์ของเว็บเพจ โดยเพิ่มเติมข้อบังคับให้แก่เมตริกซ์ความสัมพันธ์ M_e นั่นคือเมื่อพิจารณาในกรณีที่เว็บเพจใดๆไม่มีลิงค์ชี้ออกจากเว็บเพจจะทำการสร้างลิงค์เสมือนชี้ไปยังทุกๆเว็บเพจ เพื่อให้สมการการคำนวณเพจแร็งค์ (7) สอดคล้องกลับคุณสมบัติของมาคอร์ฟ

กำหนดให้ $\vec{d}$ คือเวกเตอร์ ขนาด $N\times 1$ ที่แสดงถึงเว็บเพจใดๆที่ไม่มีลิงค์ชี้ออกซึ่งสามารถแสดงได้ดังสมการต่อไปนี้

$$d_i = \begin{cases} 1 & if\ \omega(i) = 0 \\ 0 & otherwise \end{cases} \tag{5}$$

จากสมการที่ (5) จะสามารถแสดงถึงเมตริกซ์ความสัมพันธ์ M ที่เป็นตัวแทนโครงสร้างความสัมพันธ์ของเว็บเพจได้ดังสมการต่อไปนี้

$$M = M_e + (\vec{d} \times \vec{E}^T)^T \tag{6}$$

จากสมการที่ (6) และเทเลพอตเวกเตอร์ $\vec{E}$ กำหนดให้ c มีค่าเท่ากับ 0.85 (ซึ่งเป็นค่าที่ได้จากการทดลอง) เราจะสามารถคำนวณเพจแร็งค์ของเว็บกราฟใดๆได้ดังสมการต่อไปนี้

$$\vec{R}_{i+1} = cM\vec{R}_i + (1-c)\vec{E} \tag{7}$$

สมการที่ (7) ข้างต้นเป็นสมการสำเร็จที่ได้จากการปรับเมตริกซ์ความสัมพันธ์ให้สอดคล้องตามคุณสมบัติที่สำคัญอันหนึ่งของมาคอร์ฟ กล่าวคือ

1. เมื่ออยู่ที่ตำแหน่งใดๆในเว็บกราฟที่แสดงด้วยเมตริกซ์ M แล้วตามลิงค์ของเว็บกราฟไปต้องสามารถไปยังทุกๆตำแหน่งในเว็บกราฟได้ (irreducible)

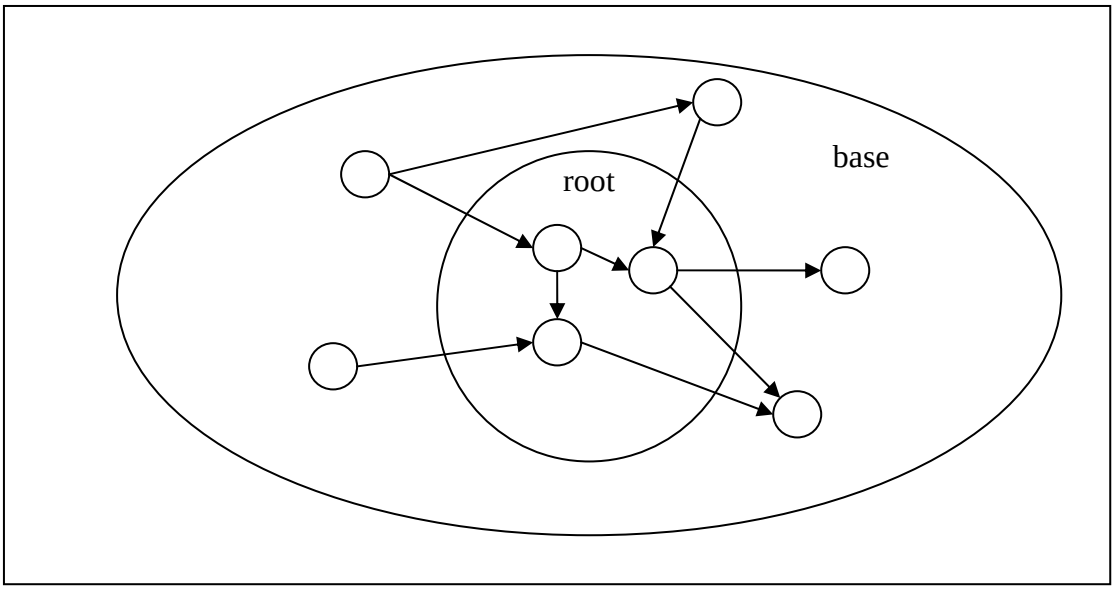
2. เมื่อตามลิงค์ของเว็บกราฟที่แสดงด้วยเมตริกซ์ M ไปยังตำแหน่งใดๆแล้วนั้นต้องสามารถมีทางย้อนกลับมายังตำแหน่งเริ่มต้นโดยใช้ระยะทาง γ, 2γ, $3\gamma, \ldots$ เมื่อกำหนดให้ $\gamma = 1 (\text{aperiodic})$

เมื่อเมตริกซ์ความสัมพันธ์มีคุณสมบัติดังกล่าวแล้วจะรับประกันได้ว่าเมื่อจำนวนรอบในการคำนวณ $i = \infty$ แล้วได้คำตอบของแร็งค์เวกเตอร์ลู่เข้าเสมอ

ฮิตส์ (HITS)

ฮิตส์เป็นเทคนิคการจัดลำดับเว็บเพจจากผลค้นคืนของระบบสืบค้นข้อมูลโดยการวิเคราะห์โครงสร้างลิงค์เช่นเดียวกับเพจแร็งค์ซึ่งถูกนำเสนอ *ใน* ค.ศ. 1999 Kleinberg โดยจุดประสงค์ต้องการแก้ปัญหาผลค้นคืนจากคำค้นหาที่จัดอยู่ประเภทคำที่มีหัวข้อที่หลากหลาย (broad topic) เช่นคำสืบค้น “Harward” ซึ่งมีเว็บเพจจำนวนมากมายที่อยู่คนละหัวข้อที่ประกอบด้วยคำๆนี้ ทำให้ผลค้นคืนของคำสืบค้นนี้จะมีปริมาณผลค้นคืนจำนวนมากเกินกว่าผู้สืบค้นสามารถตรวจสอบได้ ดังนั้นในงานวิจัยฮิตส์จึงได้นำเสนอมาตรวัดสองประเภทคือ ฮับ (hub) และ ออเทอริตี้ (authority) เพื่อใช้จัดลำดับผลค้นคืนของระบบสืบค้นข้อมูล โดยที่ให้นิยามของฮับและออเทอริตี้ไว้ดังนี้ “เว็บเพจที่เป็นฮับคือเว็บเพจที่มีเส้นทาง(ลิงค์) ไปยังเว็บเพจที่น่าเชื่อถือจำนวนมาก และเว็บเพจที่เป็นออเทอริตี้คือเว็บเพจที่มีความน่าเชื่อถือและถูกอ้างอิงจากเว็บเพจที่เป็นฮับจำนวนมาก” ดังนั้นเมื่อพิจารณาเว็บเพจใดๆจะมีค่าฮับสูงก็ต่อเมื่อเว็บเพจนั้นมีลิงค์ชี้ไปยังเว็บเพจที่มีความน่าเชื่อถือจำนวนมากและในทางกลับกันเว็บเพจใดๆนั้นจะมีค่าออเทอริตี้สูงก็ต่อเมื่อเว็บเพจนั้นถูกชี้โดยเว็บเพจที่มีลิงค์ชี้ไปยังเว็บเพจที่มีความน่าเชื่อถือจำนวนมาก จึงสามารถบอกได้ว่าเว็บเพจนั้นจะเป็นเว็บเพจที่มีความน่าเชื่อถือเช่นกัน

ในการคำนวณมาตรวัดฮับและออเทอริตี้ จำเป็นต้องสร้างเว็บกราฟที่เป็นตัวแทนของกลุ่มเว็บเพจที่ถูกค้นคืนจากคำสืบค้นใดๆ ซึ่งวิธีการนี้นั่นเองเป็นวิธีการที่เราใช้เก็บรวมรวมโครงสร้างเว็บกราฟที่ค้นคืนจาก Yahoo API ในงานวิจัยฉบับนี้



ภาพที่ 4 การสร้างเว็บกราฟที่สัมพันธ์กับคำสืบค้น

ที่มา: Kleinberg (1999)

ในขั้นตอนเริ่มต้นนั้นจะสร้างเว็บกราฟย่อยจากเว็บกราฟที่มีอยู่ทั้งหมดเพื่อนำเสนอโครงสร้างเว็บกราฟที่สัมพันธ์กับคำสืบค้นนั้นๆ โดยนำผลค้นคืนจากคำสืบค้นสร้างเป็นกลุ่มเว็บเพจเริ่มต้น (root set) หลังจากนั้นจึงขยายขนาดของกลุ่มเว็บเพจเริ่มต้นตามลิงค์ที่ชี้ออกและลิงค์ที่ชี้มายังกลุ่มของเว็บเพจเริ่มต้น ซึ่งเซตเว็บเพจที่ได้มาในขั้นตอนการขยายเว็บเพจเริ่มต้นนี้เราจะเรียกว่าเบสเซต (base set) เมื่อขยายเว็บเริ่มต้นเสร็จแล้วเราจะได้เว็บกราฟย่อยที่สัมพันธ์กับคำสืบค้น $G^* = (\mathcal{V}^*, \mathcal{E}^*)$ โดยกำหนดให้ $\mathcal{V}^*$ คือเซตของเว็บเพจในเว็บกราฟย่อย และ $\mathcal{E}^*$ คือเซตของลิงค์ที่เชื่อมต่อระหว่างเว็บเพจในเว็บกราฟดังกล่าว เราจะสามารถสร้างเมตริกซ์ความสัมพันธ์ M เป็นตัวแทนโครงสร้างความสัมพันธ์ของเว็บเพจต่างๆเหล่านั้นของเว็บกราฟย่อยได้ดังสมการต่อไปนี้

$$M(p,q) = \begin{cases} 1 & if \ (p,q) \in \mathcal{E}^* \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

กำหนดให้ $\vec{A}_i$ และ $\vec{H}_i$ เป็นเวกเตอร์ขนาด $N \times 1$ เมื่อ $N = \left|V^*\right|$ แสดงค่าออเทอริตี้และค่าฮับของเว็บเพจทั้งหมดในเว็บกราฟย่อยในรอบการคำนวณที่ i ตามลำดับ กำหนดให้ค่าฮับและออเทอ ริตี้เริ่มต้นของทุกเว็บเพจมีค่าเท่ากับ 1 ดังนั้นเราจะได้เวกเตอร์ฮับและออเทอริตี้เริ่มต้นมีค่า $\vec{A}_0 = \vec{H}_0 = [1]_{N\times 1}$ จากตัวแปรที่กำหนดให้ นักวิจัย Klienberg ได้เสนอสมการการคำนวณค่าฮับและออเทอริตี้ดังสมการต่อไปนี้

<u>มาตรวัดออเทอริตี้</u>

$$\vec{A}_{i+1} = M^T \vec{H}_i \tag{9}$$

<u>มาตรวัดฮับ</u>

$$\vec{H}_{i+1} = M\vec{A}_i \tag{10}$$

จากสมการที่ (9) และ (10) นั้นจะมีลักษณะการคำนวณแบบวนซ้ำ ที่จะคำนวณสลับไปเรื่อยๆ โดยในการคำนวณแต่ละรอบการคำนวณนั้นต้องทำการนอมอลไลท์ (normalization) เวกเตอร์ $\vec{A}_i$ และ $\vec{H}_i$ ให้ผลรวมของแต่ละเวกเตอร์มีเท่ากับ 1 เสมอ

ในทางปฏิบัติแล้ว เราจะคำนวณสมการที่ (9) และ (10) เป็นจำนวนรอบ i จนกว่าเวกเตอร์ $\vec{H}_{i+1}$ และ $\vec{A}_{i+1}$ ลู่เข้าสู่ค่าซึ่งเป็นคำตอบโดยในแต่ละรอบของการคำนวณ เราสามารถตรวจสอบการลู่เข้าของเวกเตอร์ทั้งสองได้ เช่นเดียวกันกับที่ใช้ในสมการที่ (4) กล่าวคือ ถ้าอัตราส่วนความคาดเคลื่อนของเวกเตอร์ต่อขนาดเวกเตอร์ $\vec{H}_i, \vec{H}_{i+1}$ และอัตราส่วนความคาดเคลื่อนของเวกเตอร์ต่อขนาดเวกเตอร์ $\vec{A}_i, \vec{A}_{i+1}$ มีค่าน้อยกว่าค่าที่ยอมรับได้ δ (threshold) ค่าหนึ่ง ก็จะแสดงว่าเวกเตอร์ $\vec{H}_{i+1}$ และ $\vec{A}_{i+1}$ ลู่เข้าสู่ค่าซึ่งเป็นคำตอบและหยุดการคำนวณได้

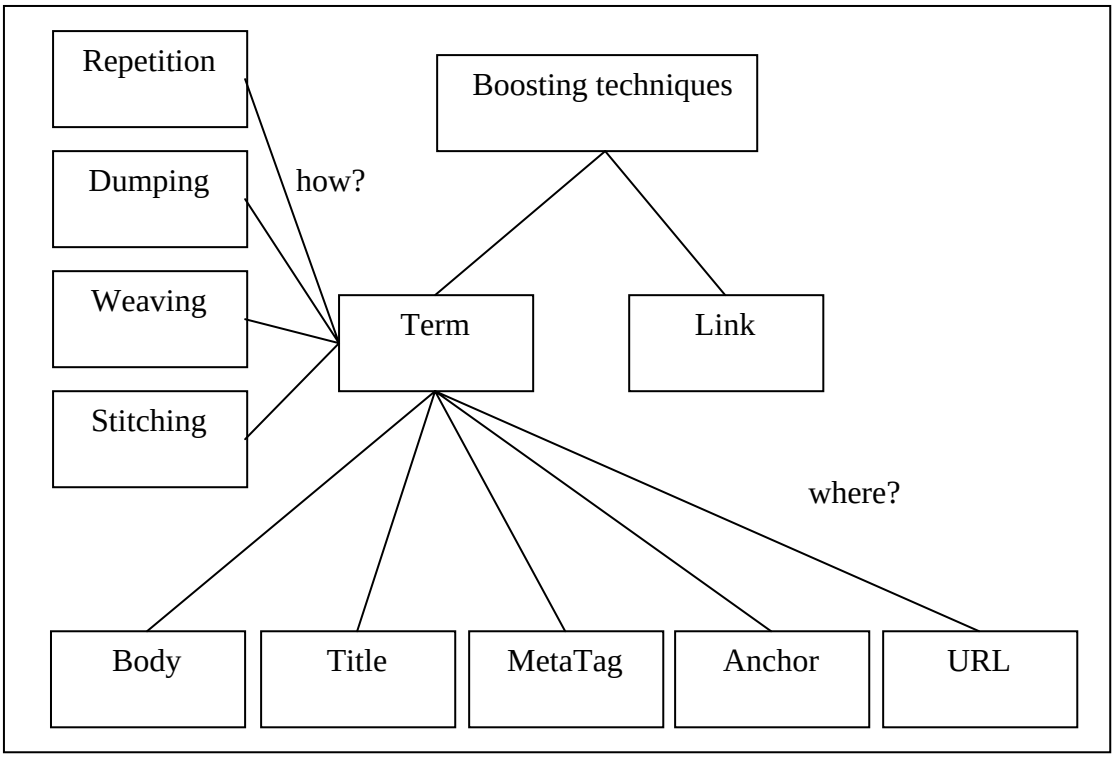
ประเภทการสแปมระบบสืบค้นข้อมูล

การสแปมระบบสืบค้นข้อมูลนั้นคือการสร้างโครงสร้างลิงค์หรือเนื้อหาของเอกสารให้ถูกคำนวณโดยระบบจัดลำดับของระบบสืบค้นข้อมูลให้ได้ค่าคะแนนในการจัดลำดับสูงๆ ทั้งที่คุณภาพของเว็บเพจนั้นไม่ได้ถูกพัฒนาตามค่าคะแนนที่เพิ่มขึ้นไปด้วย ซึ่งผลกระทบของการสแปมระบบสืบค้นข้อมูลนั้น จะทำให้ผลค้นคืนของระบบสืบค้นข้อมูลเหล่านั้นไม่น่าเชื่อถือเนื่องจากจะมีเว็บเพจที่คุณภาพต่ำปรากฏอยู่ในลำดับสูง และยังทำให้ระบบจัดลำดับของระบบสืบค้นข้อมูลต้องจัดการทำดัชนีกับเว็บเพจที่เนื้อหาไม่มีคุณภาพจำนวนมาก

รูปแบบการสแปมระบบสืบค้นข้อมูลนั้นมีได้หลากหลายเนื่องจากการสแปมระบบสืบค้นข้อมูลจะมีรูปแบบขึ้นกับวิธีการจัดลำดับของระบบสืบค้นข้อมูล ดังนั้นจึงมีนักวิจัยหลายท่านนำเสนองานวิจัยเกี่ยวกับวิธีการจัดประเภทการสแปมระบบสืบค้นข้อมูลอาทิเช่น งานวิจัยของ (Gyongyi and Garcia-Molina 2005a; Perkins, 2001) ในงานวิจัยนี้เราจะยึดตามวิธีการจัดประเภทการสแปมระบบสืบค้นข้อมูลของ Gyongyi และ Garcia ซึ่งได้แบ่งประเภทการ สแปมระบบสืบค้นข้อมูลออกเป็น 2 เทคนิคคือ เทคนิคการเพิ่มค่าคะแนนการจัดลำดับ และเทคนิคการปิดบังวิธีการทำสแปม ซึ่งเราจะกล่าวในหัวข้อย่อยถัดไปตามลำดับ

1. เทคนิคการเพิ่มค่าคะแนนการจัดลำดับ (boosting techniques)

เทคนิคการเพิ่มค่าคะแนนการจัดลำดับเป็นเทคนิคที่มุ่งเน้นทำให้เว็บเพจที่ทำการสแปมระบบสืบค้นข้อมูลนั้นได้ “ค่าความสัมพันธ์” ในตัวเนื้อหา (relevance) ซึ่งเป็นค่าที่แสดงถึงความสัมพันธ์ของคำสืบค้นกับเอกสารที่ค้นคืน ให้มีค่าความสัมพันธ์ที่สูงเมื่อสืบค้นด้วยคำสืบค้นที่เจาะจงหรือกลุ่มของคำสืบค้นที่มีความหลากหลาย และ “ค่าความสำคัญ” (importance) ของเว็บเพจเมื่อคำนึงถึงโครงสร้างการเชื่อมโยงหรือลิงค์ที่เชื่อมโยงมาจากเว็บเพจอื่นๆ ให้มีค่าความสำคัญสูงเมื่อคำนวณกับเว็บกราฟใดๆ จากที่กล่าวมานักวิจัยท่านอื่นๆ (Gyongyi and Garcia-Molina, 2005a) ได้แบ่งประเภทเทคนิคการเพิ่มค่าคะแนนการจัดลำดับดังภาพที่ 5 ออกเป็น 2 ประเภทด้วยกันคือ



ภาพที่ 5 ประเภทการสแปมเนื้อหาเอกสาร

ที่มา: Gyongyi and Garcia-Molina (2005a)

1.1 การสแปมเนื้อหาเอกสาร (term spamming)

การสแปมเนื้อหาเอกสารคือการทำการสแปมระบบสืบค้นข้อมูลโดยเพิ่มกลุ่มคำที่มีความหมายคล้ายคลึงกันหรือกลุ่มคำที่มีความหมายหลากหลายให้กับเอกสารเป้าหมาย เพื่อให้เอกสารนั้นได้ค่าคะแนนความสัมพันธ์ด้านเนื้อหาสูงเมื่อสืบค้นด้วยคำสืบค้นที่เจาะจงหรือกลุ่มของคำสืบค้นที่มีความหลากหลาย โดยสามารถจัดประเภทโดยใช้ลักษณะการเพิ่มของกลุ่มคำหรือตำแหน่งแท็ก (tag) ที่ทำการสแปมได้ดังต่อไปนี้

1.1.1 จัดประเภทโดยลักษณะการเพิ่มของกลุ่มคำ

ก. การเพิ่มคำซ้ำ (repetition) คือการเพิ่มกลุ่มคำที่มีความหมายคล้ายคลึงกันให้กับเอกสารเป้าหมาย วิธีการนี้มุ่งเน้นให้เอกสารเป้าหมายมีความสัมพันธ์สูงเฉพาะคำสืบค้นใดๆ เมื่อผู้ที่ทำการสืบค้นใช้คำสืบค้นเหล่านั้นในการสืบค้น เอกสารเป้าหมายดังกล่าวถูกค้นคืนด้วยค่า

ความสัมพันธ์ที่สูง หรืออาจจะถูกจัดลำดับในลำดับที่สูงถ้าระบบสืบค้นข้อมูลนั้นใช้ค่าความสัมพันธ์ในการจัดลำดับเป็นหลัก

ข. การเพิ่มคำซ้ำที่มีความหมายหลากหลาย (dumping) คือการเพิ่มกลุ่มคำที่มีความหมายหลากหลาย(คำที่มีในพจนานุกรม)ให้กับเอกสารที่เป็นเป้าหมาย โดยที่มุ่งเน้นให้เอกสารนั้นมีความสัมพันธ์กับคำสืบค้นทั้งหมด โดยไม่จำเป็นที่จะต้องมีความสัมพันธ์สูงเฉพาะคำสืบค้นใดๆ เมื่อผู้ที่ทำการสืบค้นใช้คำสืบค้นใดๆก็ตามที่ปรากฏในพจนานุกรม เอกสารเป้าหมายก็จะถูกค้นคืนเสมอ ซึ่งวิธีการนี้จะมีประสิทธิภาพมากสำหรับคำสืบค้นประเภท ศัพท์ทางวิชาการ หรือศัพท์ทางการแพทย์

ค. การแทรกคำ (weaving) คือการแทรกคำลงในเนื้อความของเอกสารเป้าหมายเพื่อปิดบังการทำสแปมเนื้อหาเอกสารจากสายตาของผู้ใช้งาน ซึ่งวิธีการนี้จะสามารถทำได้ทั้งรูปแบบการเพิ่มคำซ้ำหรือการเพิ่มคำซ้ำที่มีความหมายหลากหลาย โดยจะมีลักษณะดังตัวอย่างต่อไปนี้

เอกสารต้นฉบับ

Remember not only to say the right thing in the right place, but far more difficult still, to leave unsaid the wrong thing at the tempting moment.

เอกสารหลังการทำการแทรกคำ

Remember not only airfare to say the right plane tickets thing in the right place, but far cheap travel more difficult still, to leave hotel rooms unsaid the wrong thing at vacation the tempting moment.

ง. การเชื่อมต่อประโยค (phase stitching) คือการสร้างเอกสารใหม่อย่างรวดเร็วโดยใช้ประโยคของเอกสารอื่นๆ นำมารวมกันเป็นเอกสารใหม่ โดยที่เอกสารที่สร้างขึ้นใหม่โดยวิธีการนี้ไม่จำเป็นจะต้องเป็นเอกสารที่ความหมาย โดยที่จุดประสงค์ของวิธีการนี้มุ่งเน้นให้สามารถสร้างเอกสารได้อย่างรวดเร็ว และเอกสารที่สร้างขึ้นใหม่นั้นมีความสัมพันธ์กับคำสืบค้นที่

สามารถใช้สืบค้นเอกสารที่เป็นต้นฉบับได้ ลักษณะการเชื่อมประโยคจะมีลักษณะดังตัวอย่างต่อไปนี้

The objective of a search engine is to provide high-quality results by correctly identifying. Unjustifiably favorable boosting techniques, i.e., methods through which one seeks relies on the identification of some common features of spam pages.

โดยเกิดจากการนำประโยค 3 ประโยคมาสร้างเอกสารใหม่คือ

1) The objective of a search engine is to provide high quality results by correctly identifying.

2) Unjustifiably favorable boosting techniques, i.e., … และ

3) methods through which one seeks relies on the identification of some common features of spam pages.

1.1.2 จัดประเภทโดยตำแหน่งแท็กที่ทำการสแปม

ก. การสแปมเนื้อหา (body spam) ในการสแปมในส่วนเนื้อหาเอกสารเป้าหมายนั้นเป็นวิธีที่พบมากที่สุดและอาจกล่าวได้ว่าเป็นวิธีการสแปมที่เกิดมาพร้อมกับระบบสืบค้นข้อมูล ซึ่งผู้ทำการสแปมจะเพิ่มคำด้วยวิธีการต่างๆลงในส่วนเนื้อหาของเอกสารเป้าหมาย

ข. การสแปมหัวเรื่อง (title spam) เนื่องจากการวัดความสัมพันธ์ของคำสืบค้นกับเอกสารของระบบสืบค้นข้อมูลในปัจจุบันเชื่อว่าคำที่ปรากฏในหัวเรื่องจะถูกให้น้ำหนักในการคิดค่าความสัมพันธ์มากที่สุด ดังนั้นเราจึงพบได้บ่อยว่าจะเกิดการสแปมเนื้อหาเอกสารบริเวณหัวเรื่อง

ค. การสแปมแท็กพิเศษ (meta tag spam) แท็กพิเศษในเอกสารเฮสทีเอ็มแอล (HTML) นั้นจะปรากฏบริเวณบนสุดของเอกสารซึ่งมักเป็นเป้าหมายของการทำสแปมเนื้อหาเอกสาร ดังนั้นระบบสืบค้นข้อมูลในปัจจุบันมักให้ค่าน้ำหนักในการคำนวณค่าความสัมพันธ์มีค่า

น้อยหรือตัดการคิดค่าความสัมพันธ์จากบริเวณแท็กพิเศษเหล่านั้น ลักษณะการสแปมแท็กพิเศษมีลักษณะการสแปมเนื้อหาเอกสารดังตัวอย่างต่อไปนี้

```
<meta name="keywords" content="buy, cheap,
cameras, lens, accessories, nikon, canon">
```

ง. การสแปมคำที่เป็นลิงค์ (anchor text spam) ระบบสืบค้นข้อมูลมักจะให้น้ำหนักสูงสำหรับคำที่เป็นลิงค์ที่มีความสัมพันธ์กับเนื้อหาเอกสารที่ลิงค์ไปถึง ซึ่งการสแปมลักษณะนี้จะส่งผลกระทบทั้งเว็บเพจทั้งที่เป็นต้นทางและเว็บเพจที่เป็นปลายทางด้วยเช่นกัน ดังนั้นบ่อยครั้งจึงพบว่ามีการทำสแปมคำที่เป็นลิงค์เกิดขึ้น ตัวอย่างเช่น

```
<a href="www.spampage.net">spam, linkfarm, boost farm, pagerank,
keyword, anchor text </a>
```

จ. การสแปมยูอาร์แอล (URL spam) ระบบสืบค้นข้อมูลบางระบบนั้นอาจทำการแบ่งยูอาร์แอลออกเป็นเซตของคำเพื่อใช้ในการคำนวณค่าความสัมพันธ์ของเว็บเพจกับคำที่สืบค้น ดังนั้นผู้พัฒนาเว็บเพจบางคนอาจจะทำการสแปมยูอาร์แอลโดยการเพิ่มคำที่ทำการสแปมเนื้อหาเอกสารหรือตั้งโดเมนเนมเซอเวอร์ (DNS Server) เพื่อแปลงยูอาร์แอลที่ทำการสแปมเป็นไอพี (IP) ของเว็บเพจที่ต้องการ ลักษณะการสแปมแท็กพิเศษมีลักษณะการสแปมเนื้อหาเอกสารดังตัวอย่างเช่น

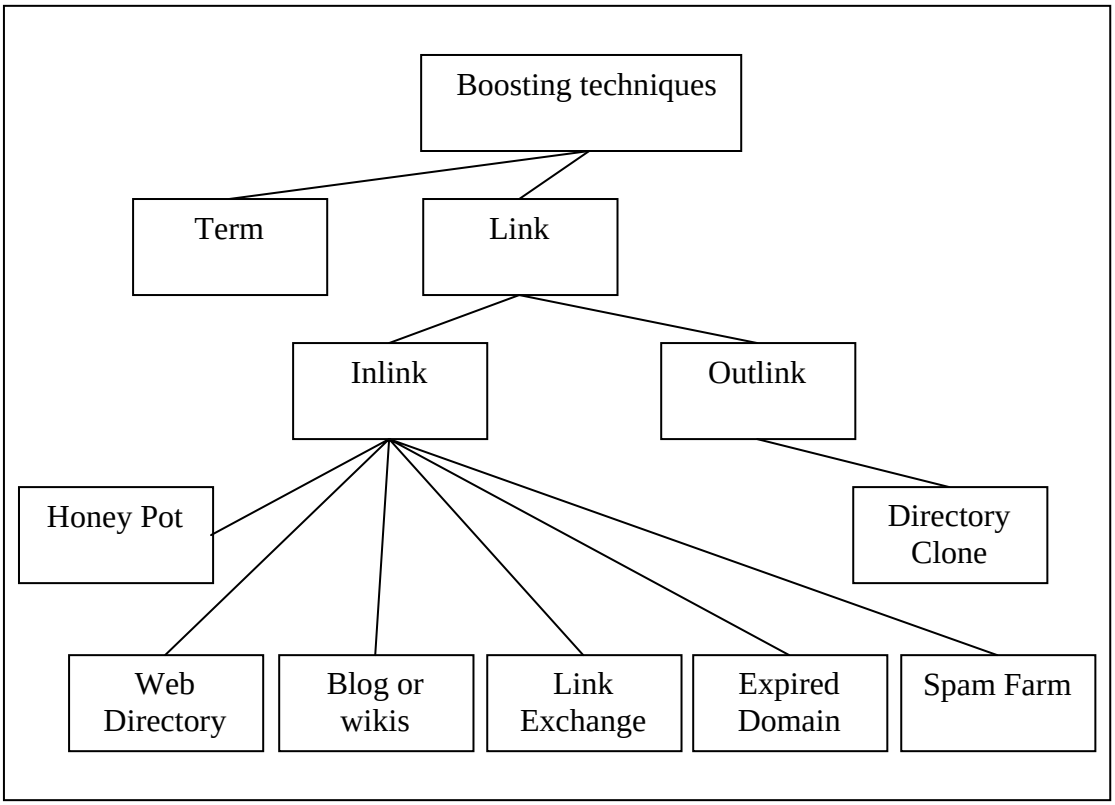
```
buy-canon-rebel-20d-lens-case.camerasx.com
buy-nikon-d100-d70-lens-case.camerasx.com
```

1.2 การสแปมโครงสร้างลิงค์ (link spamming)

การสแปมโครงสร้างลิงค์คือการทำการสแปมระบบสืบค้นข้อมูลโดยการสร้างโครงสร้างลิงค์ของกลุ่มของเว็บเพจให้ถูกคำนวณโดยระบบจัดลำดับของระบบสืบค้นข้อมูลให้ได้ค่าความสำคัญที่สูง ทั้งที่คุณภาพของเว็บเพจนั้นไม่ได้ถูกพัฒนาตามค่าคะแนนที่เพิ่มขึ้นไปด้วย ประเภทของการสแปมโครงสร้างลิงค์สามารถแยกย่อยดังภาพที่ 6 โดยใช้ลักษณะของลิงค์ และยังสามารถจัดประเภทการสแปมโครงสร้างลิงค์ตามวิธีการจัดลำดับของระบบสืบค้นข้อมูลดังต่อไปนี้

1.2.1 จัดกลุ่มโดยลักษณะลิงค์

ก. พิจารณาตามลิงค์ที่ชี้ออก (outgoing links) การสแปมโครงสร้างลิงค์ในลักษณะนี้นั้น ผู้ทำการสแปมนั้นอาจสร้างลิงค์ชี้ออกไปยังเว็บเพจที่มีคุณภาพจำนวนมาก ซึ่งการค้นหาเว็บเพจที่มีคุณภาพเหล่านี้ ผู้ทำการสแปมสามารถหาได้จากเว็บไดเร็กทอรี่ (web directory) ซึ่งเว็บไดเร็กทอรี่นั้นจะมีการจัดหมวดหมู่ของเว็บไซต์ไว้จึงง่ายต่อการทำสแปมโครงสร้างลิงค์ประเภทไดเร็กทอรี่โคลนนิ่ง (directory cloning) โดยการสร้างลิงค์ชี้ไปยังทุกๆเว็บเพจในหมวดหมู่นั้นๆที่ปรากฏในเว็บไดเร็กทอรี่



ภาพที่ 6 ประเภทการสแปมโครงสร้างลิงค์

ที่มา: (Gyongyi and Garcia-Molina, 2005a)

พิจารณาตามลิงค์ที่ชี้เข้า (incoming links) ในการสแปมโครงสร้างลิงค์เพื่อที่จะสามารถรวบรวมลิงค์ให้ชี้มายังเว็บเพจที่ต้องการนั้นผู้ที่ทำการสแปมอาจนำวิธีรวบรวมลิงค์ที่ชี้เข้าดังต่อไปนี้

1) Honey pot เป็นวิธีการสร้างเว็บเพจที่มีเนื้อหาน่าสนใจเพื่อให้ผู้พัฒนาเว็บเพจอื่นๆสร้างลิงค์ชี้มายังเว็บเพจนี้ ซึ่งภายในเว็บเพจดังกล่าวจะมีการสร้างลิงค์ชี้ไปยังเว็บเพจเป้าหมายที่ต้องการ

2) สร้างลิงค์จากบล้อคหรือวิกิพีเดีย (blog or wikis) คือการสร้างลิงค์ชี้ไปยังเว็บเพจเป้าหมาย จากบล้อคหรือวิกิพีเดียที่ผู้ดูแลไม่ได้หมั่นตรวจสอบ หรือถ้าในกรณีที่ผู้ดูแลเหล่านั้นหมั่นตรวจสอบก็ยังสามารถใช้วิธีการปิดบัง ยกตัวอย่างเช่น

Nice story. Read about my Las Vegas casinotrip.

3) แลกเปลี่ยนลิงค์ (link exchange) ผู้พัฒนากลุ่มเว็บเพจที่ทำสแปมโครงสร้างลิงค์มักจะทำการแลกเปลี่ยนลิงค์เพื่อให้โครงสร้างลิงค์ของแต่ละกลุ่มของเว็บเพจที่ทำการสแปมโครงสร้างลิงค์เชื่อมโยงกัน และเพิ่มค่าคะแนนการจัดลำดับแก่กลุ่มของเว็บเพจเป้าหมายที่ใช้วิธีการนี้ แต่อย่างไรก็ตามการทำให้แต่ละกลุ่มของเว็บเพจที่ทำการสแปมโครงสร้างลิงค์เชื่อมโยงกันอาจทำให้เว็บเพจเป้าหมายของกลุ่มของเว็บเพจที่ทำการสแปมโครงสร้างลิงค์ได้ค่าคะแนนการจัดลำดับลดลงเพื่อให้เว็บเพจเป้าหมายของกลุ่มของเว็บเพจที่ทำการสแปมโครงสร้างลิงค์อื่นๆได้ค่าคะแนนการจัดลำดับเพิ่มขึ้น

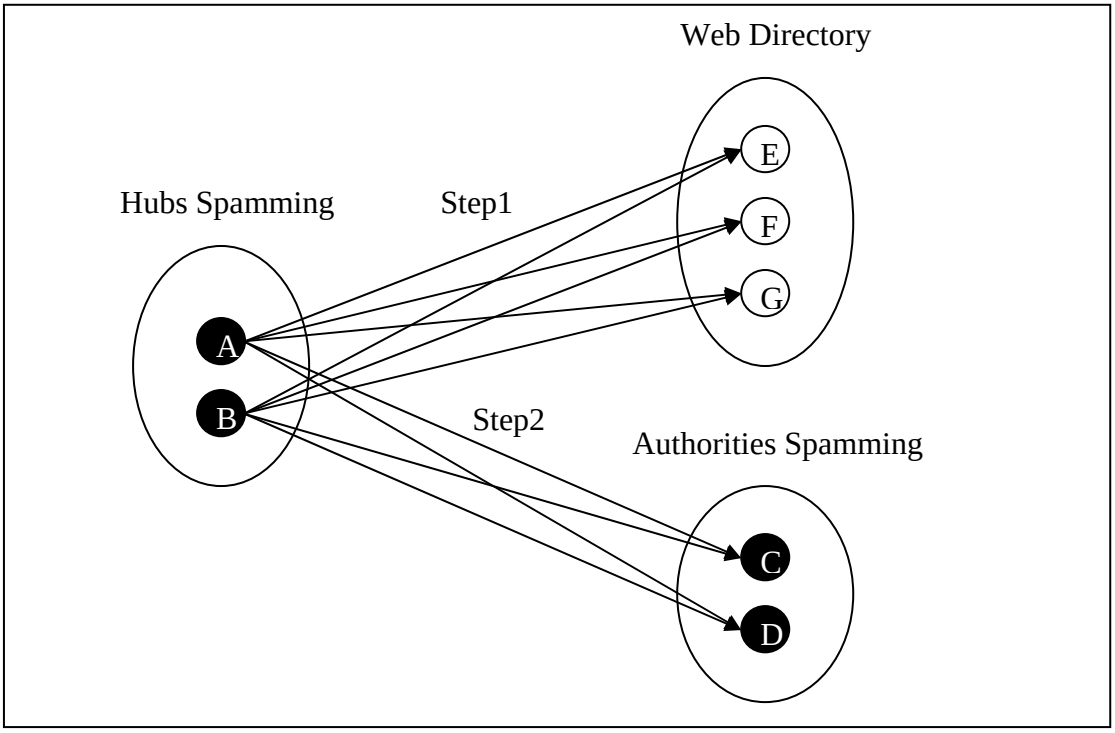
4) ซื้อชื่อโดเมนที่หมดอายุ (expired domain) เมื่อชื่อโดเมนหมดอายุลงแต่โครงสร้างลิงค์จากเว็บเพจที่ผู้ดูแลนั้นไม่ได้หมั่นตรวจสอบอาจจะยังคงมีลิงค์ชี้มายังชื่อโดเมนที่หมดอายุลง นั้น ทำให้ผู้ที่ทำการสแปมอาจจะซื้อชื่อโดเมนนั้นมาและสร้างเว็บไซต์ที่ใช้ชื่อโดเมนที่หมดอายุนั้นเพื่อนำโครงสร้างลิงค์ของชื่อโดเมนเหล่านั้นมาเพิ่มค่าคะแนนความสำคัญในการวิเคราะห์ลิงค์ให้กับเว็บเพจเป้าหมายได้

5) สร้างลิงค์จากเว็บไดเร็กทอรี่ (web directory) คือการสร้างลิงค์ชี้ไปยังเว็บเพจเป้าหมาย จากเว็บไดเร็กทอรี่ที่ผู้ดูแลเว็บไดเร็กทอรี่อนุญาตให้สร้างลิงค์ไปยังเว็บเพจ และไม่ได้ตรวจสอบลิงค์อย่างเข้มงวด ซึ่งวิธีการนี้ยังผลให้ค่าฮับและเพจแร็งค์มีแนวโน้มเพิ่มขึ้น

6) สร้างสแปมฟาร์มขึ้นเอง (spam farm) วิธีการนี้ผู้ที่ทำการสแปมต้องมีเว็บไซต์ในการดูแลเป็นจำนวนมาก และสามารถใช้เว็บไซต์เหล่านั้นสร้างโครงสร้างลิงค์ที่สามารถคำนวณโดยระบบจัดลำดับของระบบสืบค้นข้อมูลให้ได้คะแนนที่สูง

1.2.2 จัดกลุ่มโดยวิธีการจัดลำดับของระบบสืบค้นข้อมูล

ก. การสแปมโครงสร้างลิงค์ที่ใช้วิธีการจัดลำดับด้วยวิธีการฮิตส์

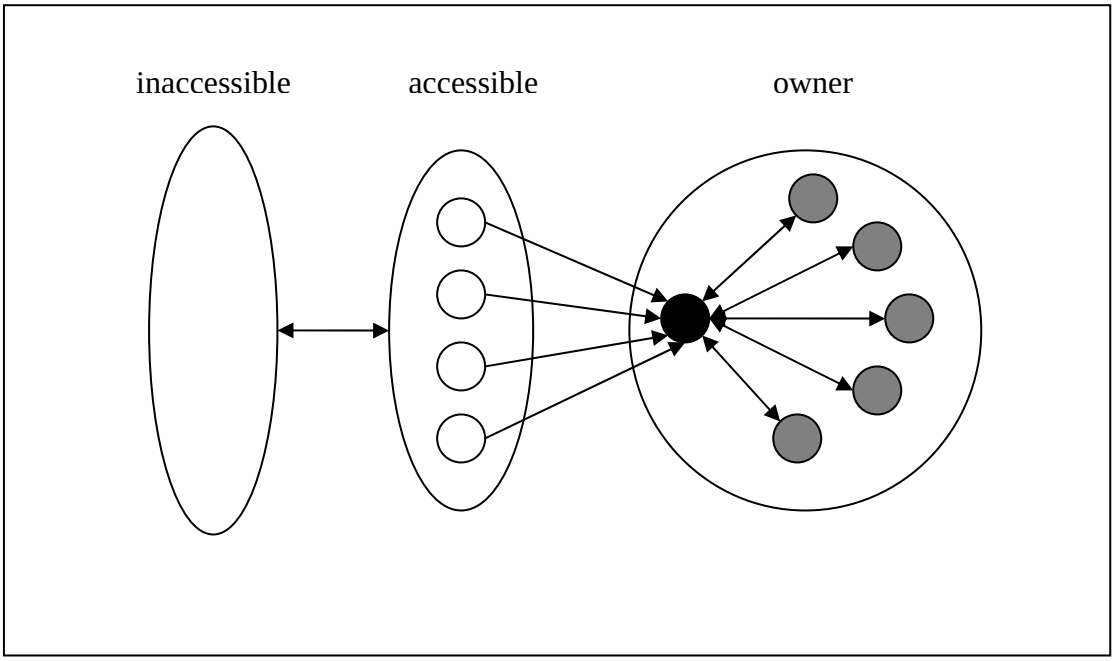


ภาพที่ 7 ตัวอย่างการสแปมการจัดลำดับวิธีการฮิตส์

วิธีการสแปมวิธีการจัดลำดับโดยวิธีการฮิตส์จากภาพที่ 7 โดยมีสมมติฐานที่ว่าเว็บเพจในเว็บไดเร็กทอรี่มีค่าออเทอริตี้ที่สูง ดังนั้นผู้ที่ทำการสแปมจะสร้างเว็บเพจ ขึ้นมากลุ่มหนึ่ง(A,B) เพื่อทำการสแปมให้เว็บเพจกลุ่มนั้นมีค่าฮับที่สูง โดยสร้างลิงค์ไปยังเว็บเพจในเว็บไดเร็กทอรี่ (E,F,G) ที่อยู่ในหมวดหมู่เดียวกัน เพื่อให้เว็บเพจ A และ Bที่สร้างขึ้นมานั้นมีค่าฮับสูง หลังจากนั้นจึงสร้างกลุ่มเว็บเพจที่ 2 (C,D) ขึ้นมา เพื่อสแปมให้เว็บเพจ C และ D นั้นเป็นเว็บเพจที่มีค่าออเทอริตี้ที่สูง จากการสร้างลิงค์ชี้มาจากกลุ่มเว็บเพจที่ทำสแปมค่าฮับมายังทุกเว็บเพจที่ต้องการทำ สแปมค่าออเทอริตี้ เมื่อทำถึงขั้นตอนนี้แล้วผู้ทำสแปมสืบค้นข้อมูลจะได้กลุ่มเว็บเพจที่มีค่าออเทอ ริตี้ที่สูงและกลุ่มของเว็บเพจที่มีค่าฮับที่สูงซึ่งสามารถนำไปใช้การสแปมโครงสร้างลิงค์สำหรับวิธีการจัดลำดับฮิตส์ แต่ในปัจจุบันเชื่อว่าวิธีการนี้ไม่ได้ถูกนำมาใช้งานแล้วเนื่องจากมีงานวิจัยมากมาย(Bharat and Henzinger, 1998; Wu and Davidson 2005b) ที่พยายามแก้ไขปัญหานี้ และการ

จัดลำดับโดยวิธีการฮิตส์นั้นเชื่อว่าไม่เป็นที่นิยมใช้งานโดยระบบสืบค้นข้อมูลเนื่องจากจะทำการจัดลำดับทุกครั้งที่มีการสืบค้นด้วยคำสืบค้นใหม่

ข. การสแปมการจัดลำดับโดยวิธีการเพจแร็งค์



ภาพที่ 8 ตัวอย่างการสแปมการจัดลำดับวิธีการเพจแร็งค์

ที่มา: Gyongyi and Garcia-Molina (2005a)

ในการสแปมค่าเพจแร็งค์ของของเว็บเพจใดๆนั้นเราสามารถสรุปวิธีการได้โดยแบ่งกลุ่มของเว็บเพจในอินเทอร์เน็ตออกเป็น 3 ประเภทดังภาพที่ 8 คือ

1) กลุ่มของเว็บเพจที่ผู้ทำการสแปมไม่สามารถเปลี่ยนแปลงโครงสร้างลิงค์ได้ (inaccessible)

2) กลุ่มของเว็บเพจที่ผู้ทำการสแปมสามารถแก้ไขหรือเปลี่ยนแปลงโครงสร้างลิงค์ได้เพียงบางส่วน เช่นเว็บเพจประเภทบล็อกหรือวิกิพีเดียเป็นต้น (accessible)

3) กลุ่มของเว็บเพจที่ผู้ทำการสแปมสามารถปรับปรุงหรือเปลี่ยนแปลงโครงสร้างลิงค์อย่างไรก็ได้ เช่น สแปมฟาร์ม เป็นต้น (owner)

ในวิธีการจัดลำดับโดยวิธีการเพจแร็งค์นั้น ผู้ที่ทำการสแปมจะเริ่มพยายามค้นหาเว็บเพจที่ผู้ที่ทำการสแปมจากส่วนที่สามารถเปลี่ยนแปลงโครงสร้างลิงค์ได้เพียงบางส่วน และสร้างลิงค์ให้ชี้ไปยังสแปมฟาร์มที่ผู้ทำการสแปมเป็นเจ้าของ โดยในส่วนของสแปมฟาร์มนี้เองผู้ที่ทำการสแปมจะสร้างให้โครงสร้างลิงค์นั้นมีลักษณะเป็นแร็งค์ซิงค์เพื่อไม่ให้สูญเสียค่าคะแนนเพจแร็งค์จากลิงค์ที่ชี้ออกจากสแปมฟาร์มไปยังเว็บเพจอื่นๆ

2. เทคนิคการปิดบังวิธีการทำสแปม (Hiding techniques)

การปิดบังการทำสแปมเป็นเทคนิคที่ไม่มีผลกระทบต่อการจัดลำดับของระบบสืบค้นข้อมูลแต่อย่างใด แต่เป็นเทคนิคที่มีไว้เพื่อปิดบังวิธีการสแปมระบบสืบค้นข้อมูลต่างๆที่กล่าวมาแล้ว จากสายตาผู้ใช้งานที่เป็นมนุษย์หรือจากผู้ที่ตรวจสอบว่าเว็บเพจนั้นๆถูกสแปมหรือไม่ ซึ่งสามารถแบ่งออกได้เป็น 3 ลักษณะย่อยคือ

2.1 การปิดบังเนื้อหา (link spamming) (content hiding) กลุ่มคำและลิงค์ที่ทำการสแปมนั้นสามารถปิดบังจากสายตาผู้ใช้งานเมื่อถูกอ่านผ่านเว็บเบาวเซอร์ (web browser) โดยวิธีพื้นฐานที่ใช้กันอย่างแพร่หลายคือการให้สีของกลุ่มคำหรือลิงค์ให้มีสีเดียวกับพื้นหลัง ยกตัวอย่างเช่น

```
<body background="white">
<font color="white">hidden text</font>
…
</body>
```

หรือในลักษณะเดียวกันผู้ที่ทำการสแปมมักจะซ่อนลิงค์ที่ชี้ไปยังเว็บเพจที่ทำการสแปมระบบสืบค้นข้อมูลโดยการสร้างภาพที่เป็นลิงค์ให้มีขนาดเล็ก และไม่สามารถมองเห็นได้ หรือมีสีเดียวกับพื้นหลัง ยกตัวอย่างเช่น

```
<a href="target.html"><img src = "tinyimg.gif"></a>
```

ซึ่งวิธีการที่กล่าวมานั้นมีจุดประสงค์คือปิดบังการสแปมระบบสืบค้นข้อมูลจากสายตาของผู้ใช้งานที่เป็นมนุษย์ แต่ยังสามารถส่งผลต่อการจัดลำดับของระบบสืบค้นข้อมูลเนื่องจากเมื่อเว็บคราวเลอร์ (web crawler) เข้ามาเก็บข้อมูลเพื่อที่จะนำไปจัดลำดับโดยระบบจัดลำดับของระบบสืบค้นข้อมูลนั้นจะเห็นเอกสารอยู่ในรูปแบบเท็กซ์ไฟล์ (text file) ทำให้กลุ่มคำ หรือลิงค์ที่ใช้วิธีการปิดบังเนื้อหาถูกจัดลำดับโดยระบบจัดลำดับของระบบสืบค้นข้อมูลด้วย

2.2 การอำพรางตัว (cloaking) เป็นวิธีการที่เว็บเซอร์เวอร์ที่เก็บเอกสารนั้นจะเลือกส่งเอกสารโดยขึ้นกับผู้ร้องขอโดยตรวจสอบจากไอพีหรือยูซเซอร์เอเจ้น (user-agent) ในแท็กพิเศษของเอกสารประเภทเฮสทีเอ็มแอล ถ้าตรวจสอบว่าเป็นผู้ร้องขอปรกติจะส่งเอกสารที่ไม่ได้ทำการสแปมให้ แต่ถ้าเป็นเว็บคราวเลอร์ที่ใช้เก็บรวบรวมข้อมูลของระบบสืบค้นข้อมูลเว็บเซอร์เวอร์ดังกล่าวก็เลือกที่จะส่งเอกสารที่ทำสแปมระบบสืบค้นข้อมูลไปให้แทน

2.3 รีไดเร็คชั่น (redirection) เป็นวิธีการที่ปิดบังการทำสแปมระบบสืบค้นข้อมูลโดยทำ การสแปมในเว็บเพจก่อนที่ทำการรีไดเร็คชั่น หลังจากรีไดเร็คชั่นไปถึงเว็บเพจปลายทางจะเป็นเว็บเพจปกติที่ไม่มีการทำทำสแปมระบบสืบค้นข้อมูล ดังนั้นผู้ใช้งานที่เป็นมนุษย์นั้นจะได้รับเว็บเพจที่ปกติหลังจากถูกรีไดเร็คชั่น แต่เว็บคราวเลอร์นั้นจะได้เว็บเพจที่ทำการสแปมระบบสืบค้นข้อมูลก่อนการทำรีไดเร็คชั่น

ตัวอย่างต่อไปนี้คือการทำรีไดเร็คชั่นโดยใช้ตามข้อกำหนดของเอสทีเอ็มแอล(HTML)

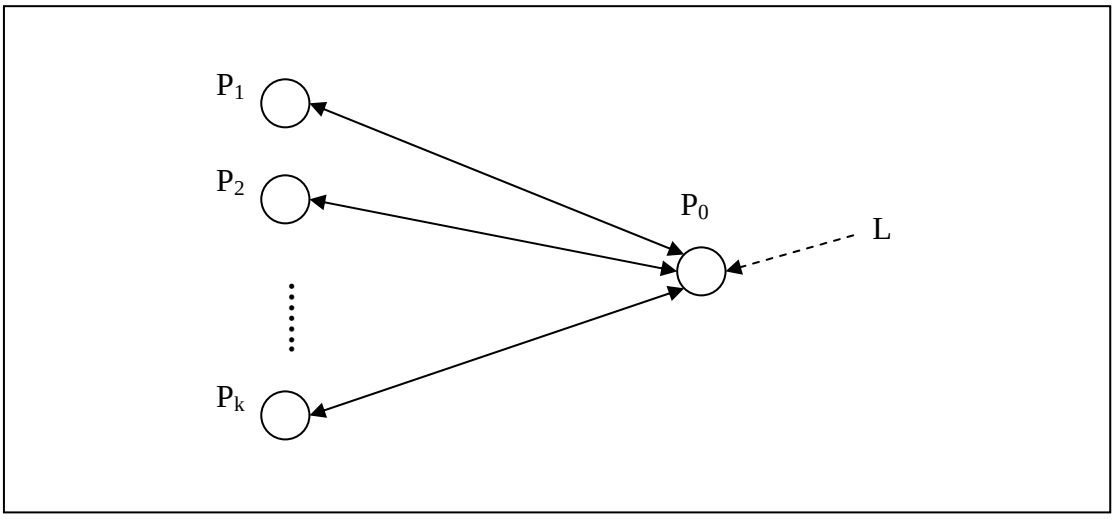
```
<html><head>
<meta http-equiv="refresh" content="0; url=http://www.example.com/">
</head><body>
 Please follow <a href="http://www.example.com/">link</a>!
<font color="white">spam spam spam spam …</font>
</body></html>
```

และ

```
HTTP/1.1 200 ok
Refresh: 0; url=http://www.example.com/
Content-type: text/html
Content-length: 78
```

ลักษณะโครงสร้างของลิงค์ที่สามารถสแปมระบบสืบค้นข้อมูล

การจัดประเภทโครงสร้างลิงค์ และผลกระทบจากการสร้างโครงสร้างลิงค์ในลักษณะต่างๆ ของการสแปมโครงสร้างลิงค์ถูกนำเสนอโดย (Gyongyi and Garcia-Molina, 2005b) โดยนักวิจัยทั้งสองได้กล่าวว่าปัญหาที่พบจากการสแปมโครงสร้างลิงค์นั้นนอกจากการที่โครงสร้างเหล่านั้นสามารถถูกทำให้คำนวณโดยระบบจัดลำดับของระบบสืบค้นข้อมูลได้คะแนนที่สูงแล้วยังยากต่อการตรวจสอบหรือค้นหา เนื่องจากมีโครงสร้างได้หลากหลายรูปแบบ แต่ทุกๆรูปแบบนั้นจะประกอบด้วยเว็บเพจสองประเภทคือ เว็บเพจช่วยเหลือ(boosting page) และเว็บเพจเป้าหมาย (target page) โดยที่เว็บเพจช่วยเหลือนั้นจะมีหน้าที่เพิ่มค่าคะแนนเพจแร็งค์ให้แก่เว็บเพจที่เป็นเป้าหมาย และเว็บเพจที่เป็นเป้าหมายนั้นจะเป็นเว็บเพจที่ผู้ที่ทำการสแปมต้องการให้มีค่าคะแนนเพจแร็งค์มากที่สุด



ภาพที่ 9 สแปมฟาร์มแบบ 1 กลุ่ม

ที่มา: Gyongyi and Garcia-Molina (2005b)

การสแปมโครงสร้างลิงค์นั้นอาจเกิดจากการรวมตัวของการสแปมโครงสร้างลิงค์ย่อยๆหลายๆกลุ่มจนเป็นกลุ่มของการสแปมโครงสร้างลิงค์ขนาดใหญ่ ซึ่งจะสามารถแบ่งออกเป็นสองลักษณะใหญ่ๆ คือ

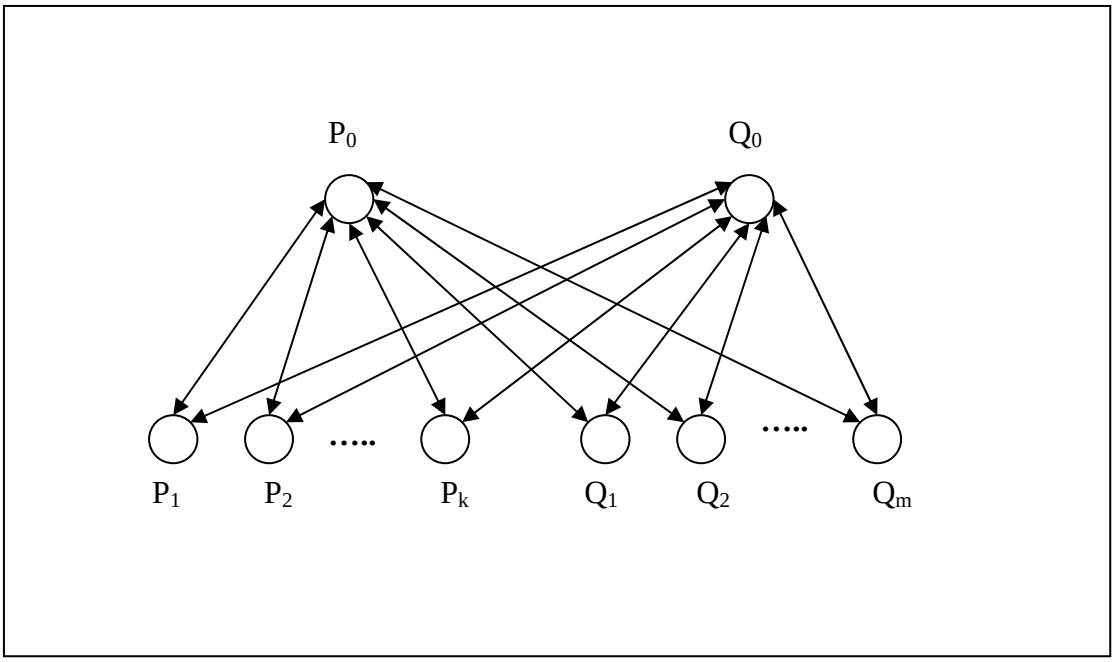
1. สแปมฟาร์มแบบ 1 กลุ่ม

สแปมฟาร์มแบบ 1 กลุ่มมีสมมติฐานว่าประกอบด้วยเว็บเพจเป้าหมาย 1 เว็บเพจและเว็บเพจช่วยเหลือในจำนวนจำกัดเนื่องจากค่าใช้จ่ายในการสร้างและดูแลเว็บเพจช่วยเหลือ โดยผู้ที่ทำการสแปมจะพยายามรวบรวมลิงค์จากเว็บเพจที่ผู้ที่ทำการสแปมสามารถเข้าไปเปลี่ยนแปลงโครงสร้างลิงค์ได้บางส่วน เช่น เว็บบอร์ดที่ผู้ดูแลไม่หมั่นตรวจสอบ ให้มีลิงค์ชี้ไปยังเว็บเพจในสแปมฟาร์ม ซึ่งรูปแบบที่จะทำให้เว็บเพจเป็นเป้าหมายได้ค่าคะแนนเพจแร็งค์สูงที่สุดเมื่อมีเว็บเพจช่วยเหลือ k เว็บเพจนั้น จะมีโครงสร้างตามภาพที่ 9 โดยกำหนดให้เว็บเพจ p1 ถึง pk จะเป็นเว็บเพจช่วยเหลือเว็บเพจ p0 เป็นเว็บเพจเป้าหมายและ L เป็นลิงค์ที่ผู้ที่ทำการสแปมรวบรวมจากเว็บเพจภายนอกสแปมฟาร์ม

2. กลุ่มรวมของสแปมฟาร์ม

กลุ่มรวมของสแปมฟาร์มมีสมมติฐานว่าเกิดจากการรวมกันของสแปมฟาร์มแบบ 1 กลุ่มหลายๆกลุ่มประกอบกัน ซึ่งลักษณะที่สแปมฟาร์มแบบ 1 กลุ่มหลายๆกลุ่มมาเชื่อมต่อกันนั้นมี 5 รูปแบบคือ

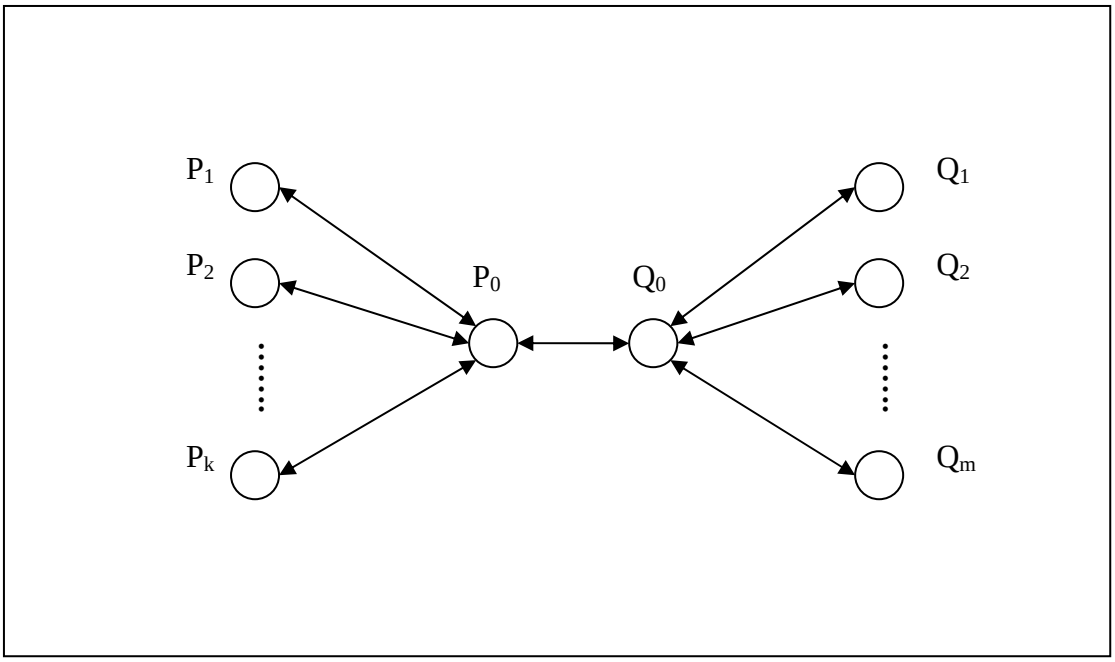
2.1 เชื่อมต่อเว็บเพจช่วยเหลือร่วมกัน จากภาพที่ 10 ประกอบด้วยสแปมฟาร์มสองกลุ่มคือ สแปมฟาร์ม P และสแปมฟาร์ม Q มีเว็บเพจ P0 และ Q0 คือเว็บเพจเป้าหมาย เว็บเพจ P1 ถึง Pk และ Q1 ถึง Qm คือเว็บเพจช่วยเหลือโดยสแปมฟาร์มกลุ่ม P และกลุ่ม Q นั้นจะทำการแชร์เว็บเพจช่วยเหลือร่วมกันโดยที่สแปมฟาร์มที่มีจำนวนเว็บเพจช่วยเหลือน้อยกว่านั้นจะได้ค่าคะแนนเพจแร็งค์เพิ่มขึ้นในการเชื่อมต่อลักษณะนี้ และสแปมฟาร์มที่มีจำนวนเว็บเพจช่วยเหลือมากกว่านั้นจะเสียค่าคะแนนเพจแร็งค์บางส่วนให้แก่สแปมฟาร์มที่มีจำนวนเว็บเพจช่วยเหลือน้อยกว่า



ภาพที่ 10 การรวมกลุ่มของสแปมฟาร์มโดยการแชร์เว็บเพจช่วยเหลือร่วมกัน

ที่มา: Gyongyi and Garcia-Molina (2005b)

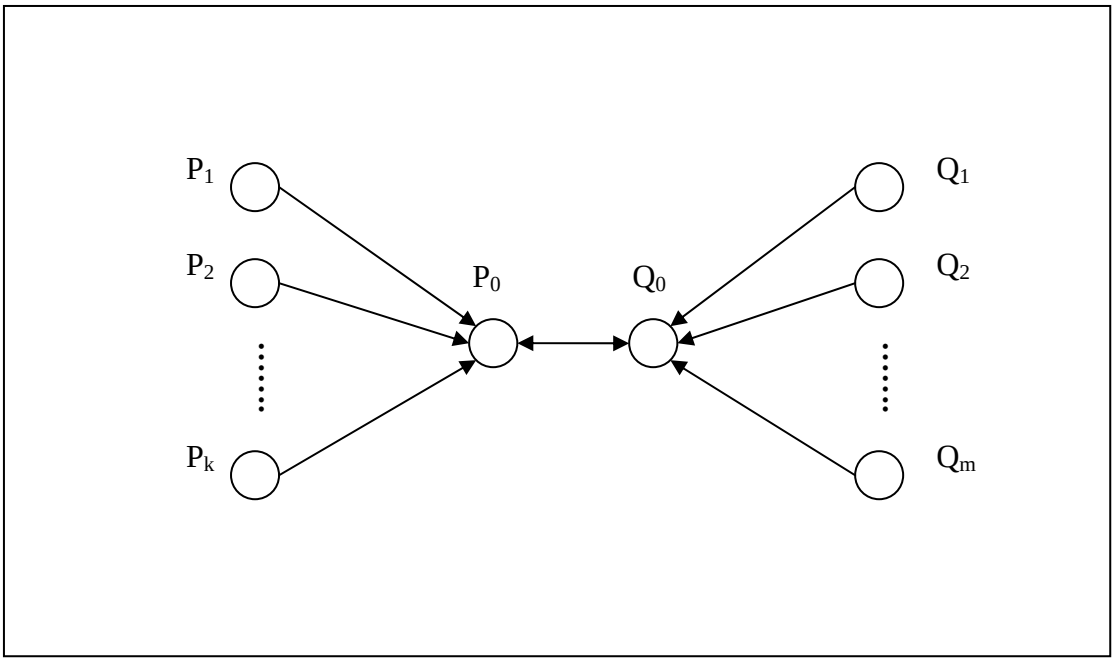
2.2 เชื่อมต่อเว็บเพจเป้าหมายร่วมกัน จากภาพที่ 11 ประกอบด้วยสแปมฟาร์มสองกลุ่มคือ สแปมฟาร์ม P และสแปมฟาร์ม Q ซึ่งมีลักษณะเว็บเพจเหมือนดังรูปแบบเดียวกันกับภาพที่ 10 นั่นคือมีเว็บเพจ P_0 และ Q_0 คือเว็บเพจเป้าหมาย เว็บเพจ P_1 ถึง P_k และ Q_1 ถึง Q_m คือเว็บเพจช่วยเหลือ แต่จะมีการเชื่อมต่อระหว่างสแปมฟาร์ม 2 กลุ่มแบบแชร์เว็บเพจเป้าหมายคือเว็บเพจ P_0 และ Q_0 เท่านั้น ซึ่งการเชื่อมต่อลักษณะนี้จะให้ผลการเพิ่มค่าคะแนนเพจแร็งค์ของเว็บเพจเป้าหมายคล้ายคลึงกับการเชื่อมต่อเว็บเพจช่วยเหลือร่วมกัน นั่นคือสแปมฟาร์มที่มีจำนวนเว็บเพจช่วยเหลือน้อยกว่านั้นจะได้ค่าคะแนนเพจแร็งค์เพิ่มขึ้นในการเชื่อมต่อลักษณะนี้และสแปมฟาร์มที่มีจำนวนเว็บเพจช่วยเหลือมากกว่านั้นจะเสียค่าคะแนนเพจแร็งค์บางส่วนให้แก่สแปมฟาร์มที่มีจำนวนเว็บเพจช่วยเหลือน้อยกว่า



ภาพที่ 11 การรวมกลุ่มของสแปมฟาร์มโดยการแชร์เว็บเพจเป้าหมายร่วมกัน

ที่มา: Gyongyi and Garcia-Molina (2005b)

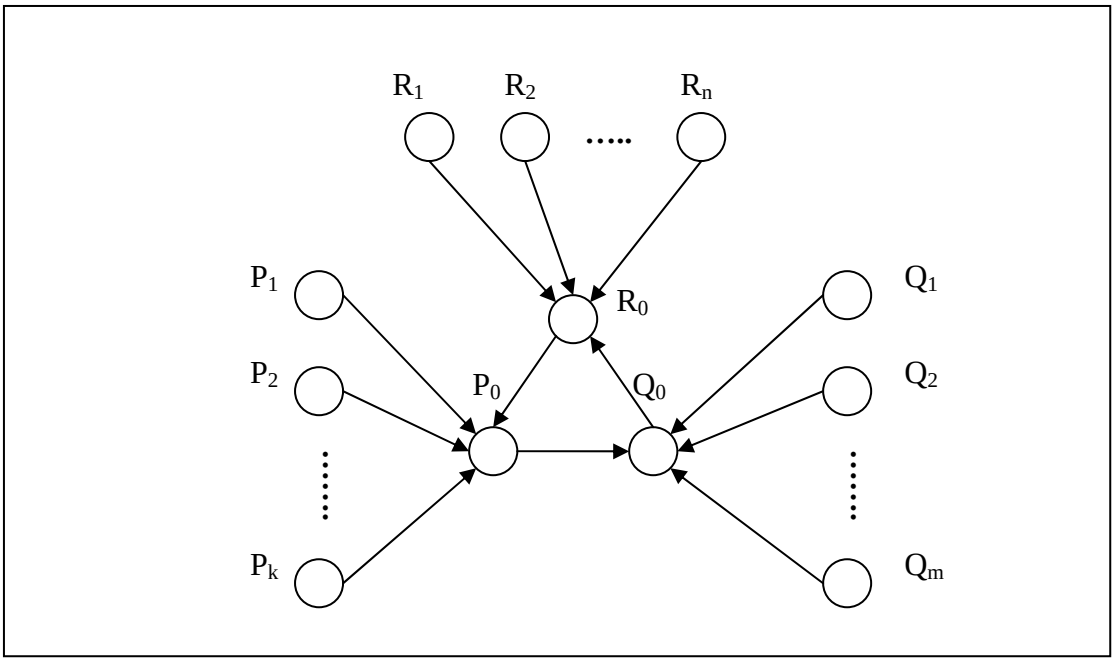
2.3 เชื่อมต่อเว็บเพจเป้าหมายร่วมกันโดยไม่มีลิงค์ชี้กลับไปยังเว็บเพจช่วยเหลือ จากภาพที่ 12 ประกอบด้วยสแปมฟาร์มสองกลุ่มคือ สแปมฟาร์ม P และสแปมฟาร์ม Q ซึ่งมีลักษณะเว็บเพจเหมือนดังรูปแบบเดียวกันกับภาพที่ 11 นั่นคือมีเว็บเพจ P_0 และ Q_0 คือเว็บเพจเป้าหมาย เว็บเพจ P_1 ถึง P_k และ Q_1 ถึง Q_m คือเว็บเพจช่วยเหลือ แต่แตกต่างกันคือจะไม่มีลิงค์ชี้กลับจากเว็บเพจเป้าหมายไปยังเว็บเพจช่วยเหลือ ซึ่งจากลักษณะการเชื่อมต่อลักษณะนี้จะทำให้กลุ่มของสแปมฟาร์มที่มาเชื่อมต่อกันได้ค่าคะแนนเพจแร็งค์ของเว็บเพจเป้าหมายเพิ่มขึ้นทั้งสองกลุ่ม โดยทั้งสองเว็บเป้าหมายนั้นจะมีค่าคะแนนเพจแร็งค์เท่ากัน



ภาพที่ 12 การรวมกลุ่มของสแปมฟาร์มโดยการแชร์เว็บเพจเป้าหมายร่วมกันโดยไม่มีลิงค์ชี้กลับไปยังเว็บเพจช่วยเหลือ

ที่มา: Gyongyi and Garcia-Molina (2005b)

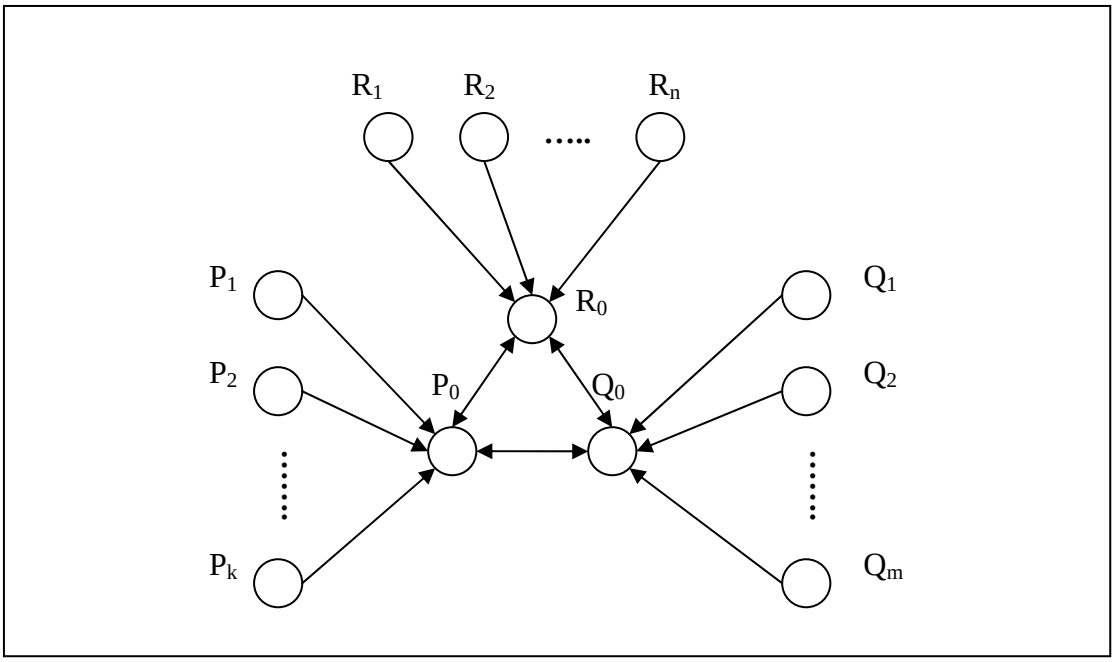
2.4 เว็บวงแหวน (web ring) จากภาพที่ 13 ประกอบด้วยสแปมฟาร์ม 3 กลุ่มคือ สแปมฟาร์ม P, Q และสแปมฟาร์ม R ซึ่งมีลักษณะเว็บเพจคือมีเว็บเพจ P_0, Q_0 และ R_0 คือเว็บเพจเป้าหมาย เว็บเพจ P_1 ถึง P_k, Q_1 ถึง Q_m และ R_1 ถึง R_n คือเว็บเพจช่วยเหลือ โดยลักษณะการเชื่อมต่อลักษณะนี้นั้นเป็นที่นิยมในหมู่ของผู้พัฒนาเว็บเพจที่มีหัวข้อสนใจในหัวข้อลักษณะเดียวกันและการเชื่อมต่อลักษณะนี้นั้นเป็นรูปแบบเริ่มต้นของโครงสร้างเว็บเพจในอินเทอร์เน็ต



ภาพที่ 13 เว็บวงแหวน

ที่มา: Gyongyi and Garcia-Molina (2005b)

2.5 เว็บโครงสร้างสมบูรณ์ (alliance with complete cores) จากภาพที่ 14 ประกอบ ด้วยสแปมฟาร์ม 3 กลุ่มคือ สแปมฟาร์ม P, Q และสแปมฟาร์ม R ซึ่งมีลักษณะเว็บเพจคือมีเว็บเพจ P_0, Q_0 และ R_0 คือเว็บเพจเป้าหมาย เว็บเพจ P_1 ถึง P_k, Q_1 ถึง Q_m และ R_1 ถึง R_n คือเว็บเพจช่วยเหลือ จากลักษณะการเชื่อมต่อลักษณะนี้จะทำให้กลุ่มของสแปมฟาร์มที่มาเชื่อมต่อกันได้ค่าคะแนนเพจแร็งค์ของเว็บเพจเป้าหมายเพิ่มขึ้นทั้งสามกลุ่ม



ภาพที่ 14 เว็บโครงสร้างสมบูรณ์

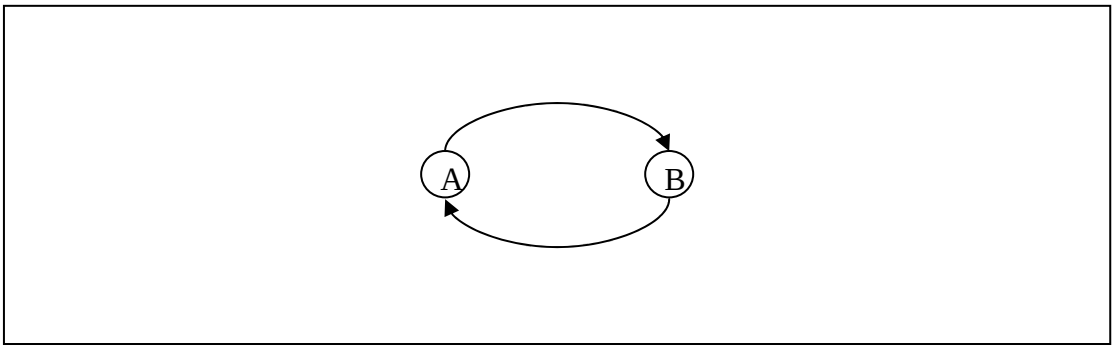
ที่มา: Gyongyi and Garcia-Molina (2005b)

วิธีการค้นหาลิงค์ฟาร์ม

ลักษณะการทำสแปมโครงสร้างลิงค์ได้ถูกนำเสนอครั้งแรกโดย (Lempel, 2000) ได้กล่าวว่าโครงสร้างลิงค์ประเภททีเคซี (tightly knit community, TKC) มีผลกระทบต่อวิธีจัดลำดับที่มีการคำนวณแบบวนซ้ำเช่นการคำนวณเพจแร็งค์และฮิตส์ หลังจากนั้นได้มีการนำเสนอแนวความคิดเรื่องลิงค์ฟาร์ม (link farm) ซึ่งเป็นรูปแบบหนึ่งของโครงสร้างลิงค์ประเภททีเคซีโดย (Wu and Brian, 2005a) โดยให้นิยามว่าเป็นเครือข่ายของเว็บไซต์ที่มีการเชื่อมโยงอย่างหนาแน่น และได้นำเสนอวิธีการค้นหาลิงค์ฟาร์มจากอินเทอร์เน็ตโดยพิจารณาเฉพาะโครงสร้างลิงค์ โดยแบ่งดังกล่าวออกเป็น 3 ขั้นตอนคือ

1. ขั้นตอนการค้นหาคอมมอนเซต (initial Step: IN-OUT common set)

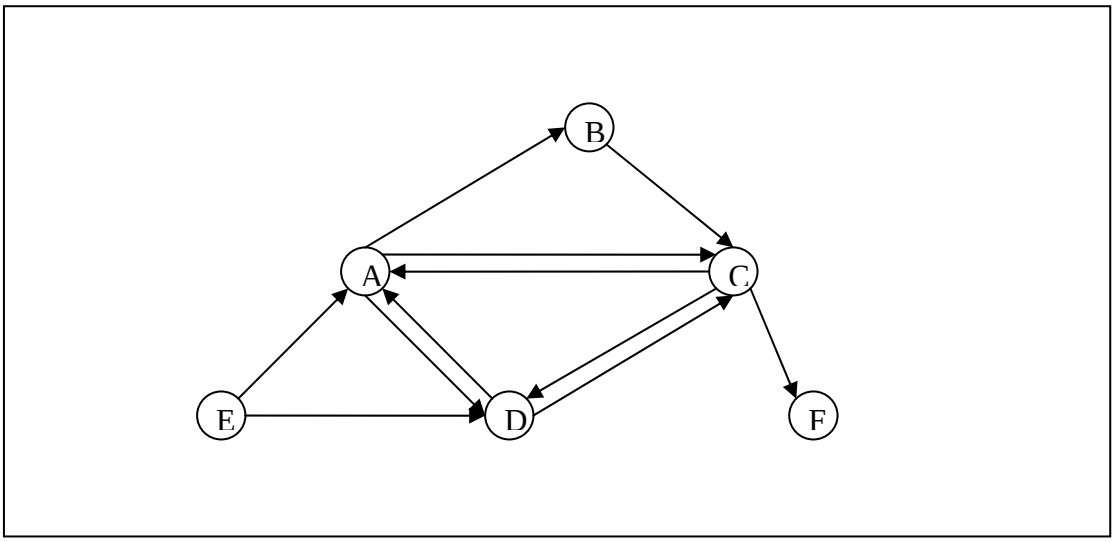
ขั้นตอนนี้เป็นขั้นตอนค้นหาเซตเริ่มต้น (seed set) จากการสังเกตนั้นผู้วิจัยได้นำเสนอว่าเว็บเพจในลิงค์ฟาร์มมักจะมีคอมมอนโหนด (common nodes) ที่มีลิงค์ชี้ไปจากเว็บเพจที่พิจารณาและมีลิงค์ชี้กลับมายังเว็บเพจที่พิจารณาดังกล่าว เราจะบอกว่าเว็บเพจที่พิจารณานั้นมีคอมมอนโหนด ดังตัวอย่างในภาพที่ 15 พิจารณาเว็บเพจ A มีคอมมอนโหนดคือเว็บเพจ B เนื่องจากเว็บเพจ A มีลิงค์ชี้ไปยังเว็บเพจ B และเว็บเพจ B มีลิงค์ชี้กลับมายังเว็บเพจ A โดยในขั้นตอนการค้นหาคอมมอนเซตนี้เองเราจะกำหนดค่าน้อยที่สุดที่ยอมรับได้ T_{io} ที่เป็นค่าที่กำหนดว่าถ้าเว็บเพจใดๆมีคอมมอนโหนดมากกว่าหรือเท่ากับค่า T_{io} เราจะเพิ่มเว็บเพจนั้นเป็นส่วนหนึ่งของเซตเริ่มต้น



ภาพที่ 15 คอมมอนโหนด

2. ขั้นตอนขยายเซตเริ่มต้น (expansion step)

ในขั้นตอนนี้มีสมมติฐานว่าเว็บเพจที่มีลิงค์ชี้ไปยังกลุ่มของเว็บเพจที่เป็นลิงค์ฟาร์ม เว็บเพจนั้นก็น่าจะเป็นส่วนหนึ่งของลิงค์ฟาร์มเช่นกัน จากสมมติฐานนี้เมื่อมีเว็บเพจใดๆมีลิงค์ชี้ไปยังเว็บเพจ ที่อยู่ในเซตเริ่มต้นเกินค่าน้อยที่สุดที่ยอมรับได้ T_{pp} ที่เรากำหนดนั้น เราจะทำการเพิ่มเว็บเพจนั้นรวมเข้ากับส่วนของเซตเริ่มต้นโดยในขั้นตอนนี้จะทำไปเรื่อยๆจนกว่าไม่มีเว็บเพจเพิ่มรวมเข้ากับเซตเริ่มต้นแล้วจึงหยุดการคำนวณ



ภาพที่ 16 เว็บกราฟตัวอย่าง

ที่มา: Wu and Brian (2005a)

จากภาพที่ 16 แสดงถึงเว็บกราฟตัวอย่าง ถ้าเรากำหนดให้ค่า $T_{io} = 2$ และค่า $T_{pp} = 2$ ในขั้นตอนเริ่มต้นเมื่อเราพิจารณาที่เว็บเพจ A นั้นจะมีคอมมอนโหนดคือเว็บเพจ C และ D ซึ่งทำให้จำนวนคอมมอนโหนดของเว็บเพจ A มีค่าเท่ากับ 2 เมื่อพิจารณาต่อมาที่เว็บเพจ C นั้นจะมีคอมมอนโหนดคือเว็บเพจ A และ D ซึ่งทำให้จำนวนคอมมอนโหนดของเว็บเพจ C มีค่าเท่ากับ 2 และเมื่อพิจารณาต่อมาที่เว็บเพจ D นั้นจะมีคอมมอนโหนดคือเว็บเพจ C และ D ซึ่งทำให้จำนวนคอมมอนโหนดของเว็บเพจ D มีค่าเท่ากับ 2 เช่นกัน ทำให้เราได้เซตเริ่มต้นจากเว็บกราฟตัวอย่างดังต่อไปนี้คือ {A, C, D} ต่อมาในขั้นตอนที่ 2 เว็บเพจ E นั้นมีลิงค์ชี้ไปยังเว็บเพจ A และ D ซึ่งเป็นเว็บเพจที่

อยู่ในเซตเริ่มต้น ดังนั้นเราจะรวมเว็บเพจ E เข้ากับเซตเริ่มต้นด้วย ซึ่งเมื่อจบขั้นตอนนี้เซตของเว็บเพจที่อยู่ในลิงค์ฟาร์มคือ {A, C, D, E}

3. ขั้นตอนจัดลำดับเว็บเพจในเซตเริ่มต้น (ranking the marked graph)

ในขั้นตอนนี้ Wu และ Brain เสนอว่าเมื่อค้นพบลิงค์ฟาร์มแล้ว เราจะมีอยู่หลายวิธีที่จะจัดการกับเว็บเหล่านั้นวิธีที่ง่ายก็คือตัดเว็บเพจเหล่านั้นออกจากการจัดลำดับ แต่พบว่าการตัดเว็บเพจเหล่านั้นออกจากการจัดลำดับนั้นอาจจะตัดเว็บเพจที่มีโครงสร้างคล้ายคลึงกับลิงค์ฟาร์มแต่เป็นเว็บเพจที่เกิดจากโครงสร้างทางธุรกิจซึ่งมีการเชื่อมโยงอย่างหนาแน่นออกดังนั้น Wu และ Brain จึงนำเสนอวิธีการ 2 วิธีการคือ

3.1 ตัดเฉพาะลิงค์ของเว็บเพจที่เป็นลิงค์ที่ชี้ออกจากเว็บเพจที่เป็นลิงค์ฟาร์ม

3.2 ปรับค่าน้ำหนักเป็นอัตราส่วนให้กับลิงค์ที่ชี้ไปจากเว็บเพจที่เป็นลิงค์ฟาร์ม

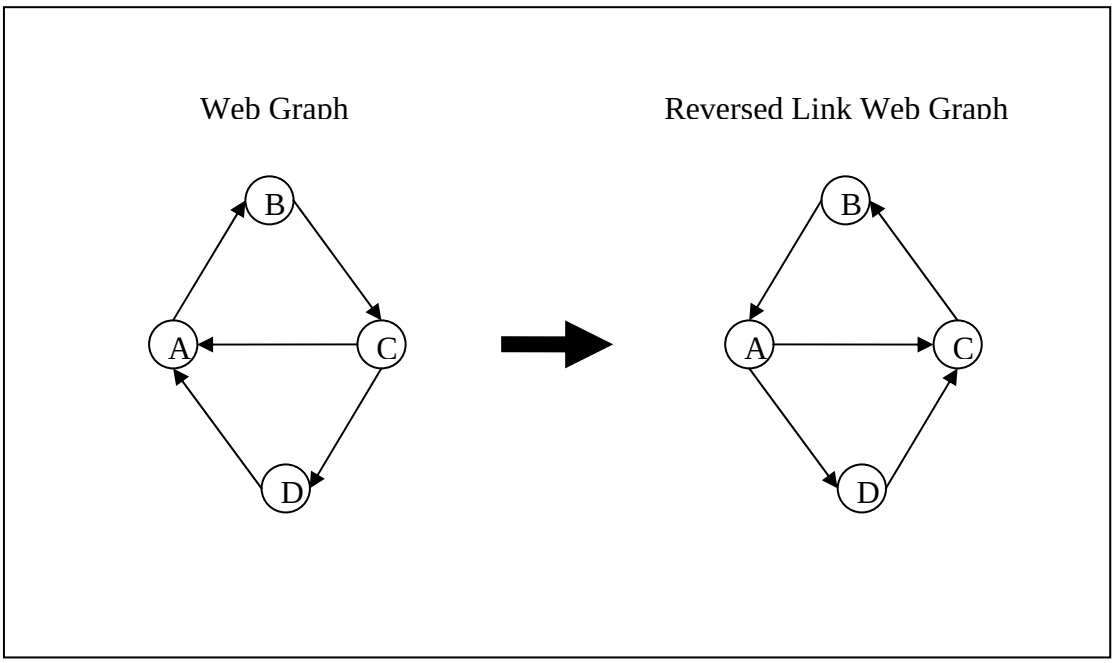
งานวิจัยที่เกี่ยวข้อง

ในหัวข้อนี้จะกล่าวถึงงานวิจัยอื่นๆที่เกี่ยวข้อง ซึ่งเป็นเทคนิคการปรับลดค่าความสำคัญของแต่ละเว็บเพจโดยนำอัลกอลิทึมเพจแร็งค์มาประยุกต์ เช่นการสร้างเทเลพอตเวกเตอร์ใหม่ซึ่งมีค่าความน่าจะเป็นของการสุ่มเดินไปยังทุกๆเว็บเพจด้วยความน่าจะเป็นไม่เท่ากันโดยพิจารณาจากคุณสมบัติของแต่ละเว็บเพจ และการคำนวณเพจแร็งค์โดยไม่พิจารณาค่าคะแนนเพจแร็งค์ของเว็บเพจในระยะทาง T โดยนำเสนอฟังค์ชั่นลดค่าพิเศษสำหรับเว็บเพจใดๆ

ทรัซแร็งค์ (trustrank)

งานวิจัยทรัซแร็งค์นั้นถูกนำเสนอ *ใน* ค.ศ. 2004 Gyongyi และ Garcia-Molina โดยมีสมมติฐานของงานวิจัยที่ว่า “เว็บเพจที่ดีนั้นจะมีลิงค์ชี้ไปยังเว็บเพจที่ดี และน้อยครั้งที่จะมีลิงค์ชี้ไปยังเว็บเพจที่ไม่ดี” โดยในงานวิจัยของทรัซแร็งค์นั้นกำหนดให้เว็บเพจที่ดีคือเว็บเพจที่ไม่ได้ถูกทำการสแปมระบบสืบค้นข้อมูล และเว็บเพจที่ไม่ดีคือเว็บเพจถูกทำการสแปมระบบสืบค้นข้อมูล ซึ่งจุดประสงค์ของงานวิจัยทรัซแร็งค์นั้นคือช่วยเหลือผู้เชี่ยวชาญในการคัดแยกเว็บเพจที่ดีออกจากเว็บเพจที่ไม่ดี โดยเริ่มจากให้ผู้เชี่ยวชาญทำการตรวจสอบเว็บเพจกลุ่มเล็กๆเพื่อสร้างเทเลพอตเวกเตอร์ $\bar{E}^*$ ที่มีขนาด $N \times 1$ ที่เบี่ยงเบนไปยังเว็บเพจที่ตรวจสอบโดยผู้เชี่ยวชาญได้ว่าเป็นเว็บเพจที่ดี และใช้เทเลพอตเวกเตอร์นั้นในการคำนวณเพจแร็งค์เพื่อคัดแยกเว็บเพจที่ดีออกจากเว็บเพจที่ไม่ดี

จากที่กล่าวมาวิธีการของทรัซแร็งค์นั้นเริ่มต้นจะกำหนดค่า L ที่แสดงถึงจำนวนเว็บเพจเริ่มต้นที่ต้องให้ผู้เชี่ยวชาญตรวจสอบ และ M_b แสดงถึงรอบการคำนวณทรัซแร็งค์ หลังจากกำหนดค่าดังกล่าวแล้ว จะทำการเลือกกลุ่มของเว็บเพจมากลุ่มหนึ่งเพื่อให้ผู้เชี่ยวชาญตรวจสอบเพื่อใช้ในการปรับค่าแก่เทเลพอตเวกเตอร์ โดยกลุ่มเว็บเพจเหล่านั้นต้องเป็นกลุ่มเว็บเพจที่เมื่อตามลิงค์ไปนั้นต้องสามารถไปถึงยังทุกๆเว็บเพจในเว็บกราฟได้หรือครอบคลุมเว็บกราฟให้มากที่สุด โดยในขั้นตอนนี้เอง นักวิจัยได้ใช้ลักษณะผลคะแนนของการคำนวณเพจแร็งค์นั่นคือ เว็บเพจที่มีค่าคะแนนเพจแร็งค์สูงคือเว็บเพจที่มีลิงค์ชี้มายังเว็บเพจนั้นจำนวนมาก ดังนั้นเมื่อทำการกลับลิงค์ทั้งหมดในเว็บกราฟแล้วทำการคำนวณเพจแร็งค์ เว็บเพจที่มีค่าคะแนนเพจแร็งค์สูงคือเว็บเพจที่มีลิงค์ชี้ออกไปยังเว็บเพจอื่นๆจำนวนมาก ทำให้เชื่อได้ว่าเมื่อตามลิงค์ของเว็บเพจเหล่านี้ไปนั้นจะสามารถครอบคลุมเว็บกราฟได้มากที่สุด ดังนั้นจึงจำเป็นต้องทำการสร้างกราฟใหม่ที่ทำการกลับลิงค์ แล้ว เพื่อใช้ในการคำนวณเพจแร็งค์ดังภาพที่ 17



ภาพที่ 17 การกลับลิงค์ของเว็บกราฟในงานวิจัยทรัซแร็งค์

หลังจากได้เว็บกราฟที่ทำการกลับลิงค์ดังกล่าวมาแล้วจะนำเว็บกราฟดังกล่าวมาสร้างเมตริกซ์ความสัมพันธ์ที่เป็นตัวแทนเว็บกราฟที่กลับลิงค์แล้ว M^* หลังจากนั้นจึงคำนวณเพจแร็งค์ตามสมการต่อไปนี้

$$\vec{R}_{i+1} = cM^*\vec{R}_i + (1-c)\vec{E} \qquad (11)$$

เมื่อได้ค่าคะแนนเพจแร็งค์ของเว็บเพจในเว็บกราฟที่กลับลิงค์แล้วจะเลือกเว็บเพจ L เว็บเพจที่มีค่าคะแนนเพจแร็งค์สูงสุดให้ผู้เชี่ยวชาญตรวจสอบซึ่งเราจะเรียกเซตของเว็บเพจที่กล่าวมาว่าเซตเริ่มต้น (seedset, S) ซึ่งในขั้นตอนนี้จะเป็นการกำหนดออราเคิลฟังค์ชั่น (oracle function, O(p)) เมื่อ p คือเว็บเพจใดๆ จะได้ฟังค์ชั่นที่เป็นตัวแทนการตรวจสอบเว็บเพจโดยผู้เชี่ยวชาญ โดยถ้าเว็บเพจที่ตรวจสอบโดยผู้เชี่ยวชาญนั้นเป็นเว็บเพจที่ดีก็จะให้ค่าเท่ากับ 1 แต่ถ้าตรวจสอบแล้วเป็นเว็บเพจที่ไม่ดีจะให้ค่าเท่ากับ 0 ซึ่งสามารถเขียนแทนด้วยสมการดังต่อไปนี้

$$O(p) = \begin{cases} 1 & \text{if } p \text{ is good} \\ 0 & \text{if } p \text{ is bad} \end{cases} \qquad (12)$$

ในขั้นตอนต่อไปจะทำการกำหนดค่าในเทเลพอตเวกเตอร์ $\vec{E}^*$ โดยถ้าเว็บเพจที่พิจารณาเป็นเว็บเพจที่เป็นสมาชิกของเซตเริ่มต้นก็จะให้ผู้เชี่ยวชาญตรวจสอบเว็บเพจนอกเหนือจากนั้นจะให้ค่าเท่ากับ 0 ทั้งหมดซึ่งสามารถเขียนแทนด้วยสมการดังต่อไปนี้

$$\vec{E}_i^* = \begin{cases} O(i) & if\ i \in S \\ 0 & otherwise \end{cases} \tag{13}$$

หลังจากขั้นตอนนี้แล้วเราจะได้เทเลพอตเวกเตอร์ที่พร้อมสำหรับการคำนวณทรัซแร็งค์ ในการคำนวณทรัซแร็งค์นั้นคือการคำนวณเพจแร็งค์ที่ใช้เทเลพอตเวกเตอร์ $\vec{E}^*$ นั่นเอง ซึ่งสามารถเขียนแทนด้วยสมการดังต่อไปนี้

$$\vec{R}_{i+1} = cM\vec{R}_i + (1-c)\vec{E}^* \tag{14}$$

โดยในที่นี้ M คือเมตริกซ์ความสัมพันธ์ของเว็บกราฟปกติ การคำนวณทรัซแร็งค์จะต้องกำหนดรอบการคำนวณ ดังนั้นจึงต้องคำนวณแบบวนซ้ำในสมการที่ (14) เป็นจำนวน M_b รอบ หลังจากนั้นจึงนำผลคะแนนทรัซแร็งค์ของแต่ละเว็บเพจมาคัดแยกเว็บเพจที่ดีออกจากเว็บเพจที่ไม่ดี โดยในขั้นตอนนี้จะกำหนดค่าน้อยที่สุดที่ยอมรับได้ (threshold) ค่าหนึ่งขึ้นมาถ้าค่าคะแนนทรัซแร็งค์ของเว็บเพจใดๆมีค่าน้อยกว่าค่าน้อยที่สุดที่ยอมรับได้เราจะคาดคะเนว่าเว็บเพจนั้นเป็นเว็บเพจที่ไม่ดีและจะให้ผู้เชี่ยวชาญตรวจสอบต่อไป

เมื่อสังเกตการณ์ทำงานของทรัซแร็งค์เป็นวิธีการค้นหาเว็บเพจที่ดีโดยการสร้างเทเลพอตเวกเตอร์ $\vec{E}^*$ ซึ่งเป็นวิธีหนึ่งในการต่อสู้กับการสแปมระบบสืบค้นข้อมูลโดยการสร้างลิงค์เสมือนเช่นเดียวกับงานวิจัยฉบับนี้ แต่มีความแตกต่างคืองานวิจัยทรัซแร็งค์นั้นจำเป็นต้องให้ผู้เชี่ยวชาญตรวจสอบเว็บกราฟเริ่มต้นและงานวิจัยทรัซแร็งค์นั้นไม่ได้มุ่งเน้นแก้ปัญหาเฉพาะการสแปมโครงสร้างลิงค์ อย่างไรก็ตามจากสมการการคำนวณทรัซแร็งค์ที่ (14) นั้นจะเห็นได้ว่าเว็บกราฟที่ใช้ในการคำนวณไม่เป็นไปตามคุณสมบัติของมาคอร์ฟ ทำให้ไม่สามารถรับประกันได้ว่าเมื่อจำนวนรอบในการคำนวณ $i = \infty$ แล้วได้คำตอบของแร็งค์เวกเตอร์ลู่เข้าสู่ค่าๆหนึ่งเสมอจึงจำเป็นต้องกำหนดรอบการคำนวณเสมอ

สแปมแร็งค์ (spamrank)

งานวิจัยสแปมแร็งค์นั้นถูกนำเสนอ *ใน* ค.ศ. 2005 Benczur และคณะ โดยมีสมมติฐานที่ว่าเว็บเพจช่วยเหลือของเว็บเพจที่ถูกสแปมโครงสร้างลิงค์นั้นจะมีการเบี่ยงเบนการกระจายตัวของค่าคะแนนเพจแร็งค์ของเว็บเพจช่วยเหลือ ยกตัวอย่างเช่นเว็บเพจที่ประกอบด้วยลิงค์ชี้มาจากเว็บเพจช่วยเหลือที่มีค่าคะแนนเพจแร็งค์ต่ำจำนวนมาก โดยงานวิจัยสแปมแร็งค์พิจารณาการกระจายตัวของค่าคะแนนเพจแร็งค์ของเว็บเพจช่วยเหลือของเว็บเพจที่ไม่ถูกสแปมโครงสร้างลิงค์นั้นจะมีการกระจายตัวแบบเพาเวอร์ลอร์ (power law distribution) โดยในงานวิจัยสแปมแร็งค์นำเสนอวิธีการสร้างเทเลพอตเวกเตอร์ *Penalty* ที่มีขนาด $N \times 1$ ที่เบี่ยงเบนไปยังเว็บเพจใดๆที่มีการกระจายตัวของค่าคะแนนเพจแร็งค์ของเว็บเพจช่วยเหลือไม่เป็นไปตามการกระจายตัวแบบเพาเวอร์ลอร์ และใช้เทเลพอตเวกเตอร์ที่กล่าวมานั้นในการคำนวณสแปมแร็งค์เพื่อนำมาปรับลดค่าคะแนนเพจแร็งค์ของเว็บเพจใดๆ

การคำนวณสมการเพจแร็งค์สามารถพิจารณาเป็นรูปแบบการสุ่มเดิน โดยที่ค่าคะแนนเพจแร็งค์ของเว็บเพจ p ใดๆนั้นหมายถึงความน่าจะเป็นที่จะสุ่มเดินมาหยุด ณ เว็บเพจ p และ ค่าคะแนนเพจแร็งค์ของเว็บเพจ p ใดๆที่เบี่ยงเบนค่าสุ่มกระโดดไปยังเว็บเพจ q ใดๆนั้นหมายถึงความน่าจะเป็นที่จะสุ่มเดินที่เริ่มต้นจากเว็บเพจ q มาหยุด ณ เว็บเพจ p โดยกำหนดให้ $PPR_q(p)$ เป็นตัวแทนการคำนวณค่าคะแนนเพจแร็งค์ของเว็บเพจ p ใดๆที่เบี่ยงเบนค่าสุ่มกระโดดในเทเลพอตเวกเตอร์ ณ ตำแหน่ง q^{th} มีค่าเท่ากับ 1 นอกจากนั้นมีค่าเป็น 0 ทั้งหมด และ c มีค่าอยู่ระหว่าง [0,1] แล้วจะสามารถแสดงการคำนวณค่าคะแนนเพจแร็งค์ของเว็บเพจ p ใดๆที่เบี่ยงเบนค่าสุ่มกระโดดไปยังเว็บเพจ q ใดๆได้ดังสมการต่อไปนี้

$$PPR_q(p) = \sum_{\substack{tour\ t: \\ q \to p}} P[t](1-c)c^{l(t)} \qquad (15)$$

เมื่อ t คือการสุ่มเดิน $(w_1,\ldots, w_k)$ ซึ่งใช้ระยะทาง $l(t) = k\text{-}1$ ซึ่งหมายถึงจำนวนลิงค์ และ $(1-c)c^{l(t)}$ คือความน่าจะเป็นการสุ่มเดิน โดยกำหนดให้ $P[t]$ คือการสุ่มเดินซึ่งมีค่าเท่ากับ $\prod_{i=1}^{k-1}\frac{1}{\omega(w_i)}$ หรือ 1 เมื่อ $l(t) = 0$ สำหรับบทพิสูจน์ได้ถูกกล่าวไว้ใน (Jed and Wisdom, 2003)

สมการการคำนวณเพจแร็งค์ที่กล่าวมาในขั้นต้นนั้นงานวิจัยสแปมแร็งค์ได้นำมาประยุกต์ใช้ในการค้นหาค่าซัพพอตที่เว็บเพจช่วยเหลือส่งมอบให้แก่เว็บเพจใดๆ และหลังจากนั้นจึงนำค่าซัพพอตเหล่านั้นมาปรับลดผลกระทบของการทำสแปมโครงสร้างลิงค์ โดยการสร้างเทเลพอตเวกเตอร์ *Penalty* ที่มีขนาด $N \times 1$ จากที่กล่าวมาทั้งหมดสามารถสรุปขั้นตอนงานวิจัยสแปมแร็งค์แบ่งออกเป็น 3 ขั้นตอนคือ

ขั้นตอนค้นหาเว็บเพจช่วยเหลือ ในขั้นตอนนี้จะคำนวณหาค่าซัพพอต ($(1-c)c^{l(t)}$) ของทุกๆเว็บเพจในเว็บกราฟโดยกำหนดขอบเขตการค้นหาเว็บเพจช่วยเหลือของเว็บเพจใดๆไว้ที่ 1000 เว็บเพจ ซึ่งสามารถเขียนรหัสเทียม (psuedo code) ในขั้นตอนนี้ได้ดังต่อไปนี้

for all web pages *i* **do**
  for *k* = 1 to 1000 **do**
    t = random value from geometric distribution with parameter $(1 - c)$
    j = endvertex of a random walk of length *t* start at *i*
    $\text{Support}_{j,i} = \text{Support}_{j,i} + (1-c)c^{l(t)}$

ขั้นตอนคำนวณค่ากระทำผิด (penalization) สำหรับทุกๆเว็บเพจในเว็บกราฟจะคำนวณค่า (correlation coefficient, ρ) ระหว่างการกระจายตัวแบบพาวเวอร์ลอร์และค่าตะแนนเพจแร็งค์ซัพพอตของเว็บเพจช่วยเหลือของเว็บเพจที่พิจาณา โดยขั้นตอนนี้จะกำหนดค่าน้อยที่สุดที่ยอมรับของค่า correlation coefficient ได้มาคือ ρ_0 ซึ่งในงานวิจัยสแปมแร็งค์กำหนดให้มีค่าเท่ากับ 0.85 ถ้าหากค่า ρ มีค่าน้อยกว่าค่า ρ_0 ก็จะนำค่า ρ และ ρ_0 มาใช้ในในการคำนวณค่ากระทำผิดในเทเลพอตเวกเตอร์ ณ ตำแหน่ง *j* ใดๆ($Penalty_j$) ดังสมการต่อไปนี้

$$Penalty_j = (\rho_0 - \rho) \times Support_{i,j} \tag{16}$$

โดยที่ในงานวิจัยสแปมแร็งค์กำหนดค่า $Penalty_j$ มีค่ามากที่สุดเท่ากับ 1 ถ้าค่า $Penalty_j$ ของเว็บเพจใดๆมีค่ามากกว่า 1 ก็จะถูกปรับค่าให้เท่ากับ 1 เสมอ

ขั้นตอนคำนวณสแปมแร็งค์ ขั้นตอนนี้คือการคำนวณเพจแร็งค์ที่เบี่ยงเบนไปยังเทเลพอตเวกเตอร์ *Penalty* ที่มีขนาดขนาด $N \times 1$ และกำหนดให้ค่าคะแนนสแปมแร็งค์ของเว็บเพจทั้งหมดในรอบการคำนวณ *i* แสดงโดยเวกเตอร์ $\vec{S}_i$ ซึ่งมีขนาด $N \times 1$ จะสามารถเขียนการคำนวณสแปมแร็งค์ดังสมการต่อไปนี้

$$\vec{S}_{i+1} = cM\vec{S}_i + (1-c) \times Penalty \quad (17)$$

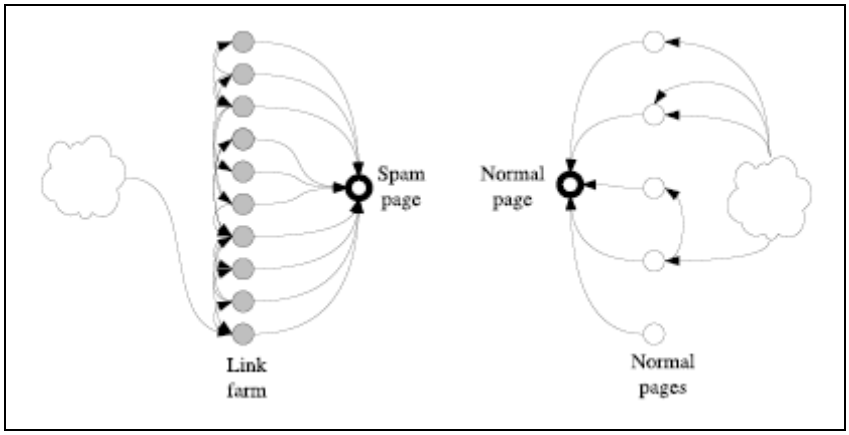
เมื่อทำงานครบทั้งสามขั้นตอนแล้วนั้นจะได้ค่าสแปมแร็งค์ของแต่ละเว็บเพจซึ่งเราจะใช้ค่านี้นั่นเองในการคำนวณค่าคะแนนที่ใช้จัดลำดับผลค้นคืนของระบบสืบค้นข้อมูล (true_authority) สามารถแสดงได้ดังสมการต่อไปนี้

$$\text{true\_authority} = \text{PageRank} - \text{SpamRank} \quad (18)$$

เมื่อสังเกตการณ์ทำงานของสแปมแร็งค์เป็นวิธีการคำนวณหาค่ากระทำผิดของแต่ละเว็บเพจโดยการสร้างเทเลพอตเวกเตอร์ *Penalty* โดยไม่ต้องการตรวจสอบจากผู้เชี่ยวชาญเหมือนงานวิจัยทรัซแร็งค์ และใช้เทเลพอตเวกเตอร์ *Penalty* ที่กล่าวมาคำนวณหาค่าสแปมแร็งค์ของแต่ละเว็บเพจเพื่อปรับลดค่าคะแนนเพจแร็งค์ เช่นเดียวกับงานวิจัยฉบับนี้ซึ่งพยายามสร้างโครงสร้างลิงค์เสมือนเพื่อปรับลดผลกระทบของการทำสแปมโครงสร้างลิงค์ อย่างไรก็ตามงานวิจัยสแปมแร็งค์นั้นจะพิจารณาปรับลดผลกระทบการทำสแปมโครงสร้างลิงค์โดยพิจารณาทีละเว็บเพจ ซึ่งในความเป็นจริงแล้วนั้นการสแปมโครงสร้างลิงค์มักเกิดจากการรวมกลุ่มของเว็บเพจเพื่อยังผลค่าคะแนนเพจแร็งค์ให้แก่เว็บเพจในกลุ่มนั้นๆ และข้อจำกัดของงานวิจัยสแปมแร็งค์นั่นคือจะไม่สามารถคำนวณอัลกอลิทึมสแปมแร็งค์กับโครงสร้างสแปมที่มีเว็บเพจช่วยเหลือจำนวนน้อยกว่าที่จะสามารถการวิเคราะห์การกระจายตัวของค่าคะแนนเพจแร็งค์ได้ ซึ่งในงานวิจัยสแปมแร็งค์นี้เมื่อเว็บเพจช่วยเหลือของเว็บเพจใดๆมีจำนวนน้อยกว่า 1000 เว็บเพจก็จะไม่พิจารณาปรับลดผลกระทบการสแปมโครงสร้างลิงค์ของเว็บเพจนั้น ซึ่งสามารถทำให้สรุปได้ว่าความถูกต้องของงานวิจัยสแปมแร็งค์จะขึ้นกับขนาดของการสแปมโครงสร้างลิงค์ และจำนวนของเว็บเพจช่วยเหลือของเว็บเพจที่พิจารณานั่นเอง

URPC อัลกอลิทึม

งานวิจัย Using Rank Propagation and Probabilistic Counting for Link Based Spam Detection (URPC) นั้นถูกนำเสนอ *ใน* ค.ศ. 2006 Becchetti และคณะ โดยกล่าวว่าเว็บเพจที่พบในลิงค์ฟาร์มนั้นมักจะมีเว็บเพจช่วยเหลือในระยะทางที่สั้นจำนวนมากแต่ในระยะระยะทางที่ไกลออกไปจะมีจำนวนน้อยกว่าควรจะเป็น ซึ่งสามารถแสดงได้ดังภาพที่ 18



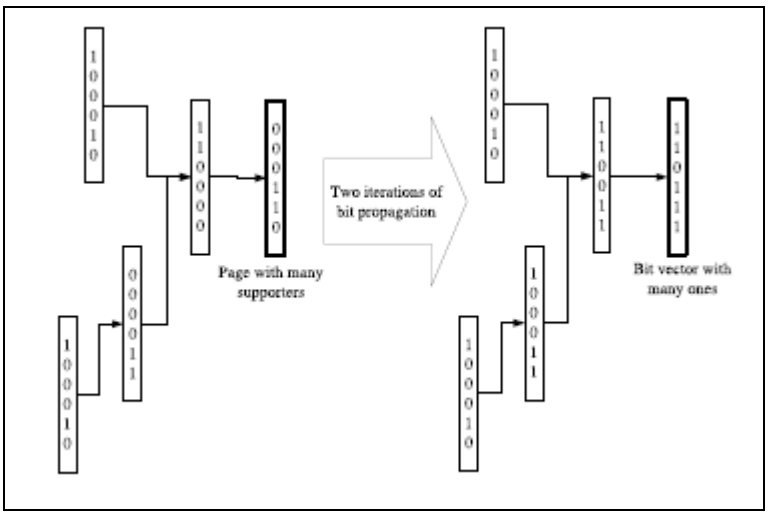
ภาพที่ 18 ลักษณะเว็บเพจช่วยเหลือของเว็บเพจที่พบในลิงค์ฟาร์ม

ที่มา: Becchetti *et al*. (2006)

และยังได้นำเสนอวิธีแก้ไขโดยใช้ วิธีการค้นหาเว็บเพจช่วยเหลือในระยะทาง *t* (neighborhood counting technique) และ วิธีการคำนวณเพจแร็งค์แบบตัดปลาย (truncated PageRank)

1. วิธีการค้นหาเว็บเพจช่วยเหลือ

เนื่องจากการค้นหาเว็บเพจช่วยเหลือทั้งหมดแก่ทุกเว็บเพจในเว็บกราฟขนาดใหญ่ เช่น อินเทอร์เน็ตนั้นไม่สามารถทำได้ งานวิจัย URPC จึงทำการค้นหาเว็บเพจช่วยเหลือมีแนวคิดพื้นฐานจากงานวิจัย (Cohen, 1997; Flajolet and Martin 1985) โดยเริ่มจากจะกำหนดบิตเวกเตอร์แก่เว็บเพจ x ใดๆ (V_x) ซึ่งมีขนาด k ซึ่งในงานวิจัย URPC ค่า k นั้นมีค่าเท่ากับ 32 และ 64 บิตโดยแต่ละสมาชิกของบิตเวกเตอร์นั้นจะมีค่าเท่ากับ 1 ด้วยความน่าจะเป็นเท่ากับ $1/N$ นอกนั้นจะให้ค่าเท่ากับ 0



ภาพที่ 19 การค้นหาเว็บเพจเป้าหมายโดยใช้บิตเวกเตอร์

ที่มา: Becchetti *et al*. (2006)

เมื่อกำหนดบิตเวกเตอร์ให้แก่ทุกเว็บเพจในเว็บกราฟแล้วนั้นจะทำการคำนวณแบบวนรอบ โดยที่เมื่อเว็บเพจ y มีลิงค์ชี้ไปยังเว็บเพจ x แล้วเว็บเพจ y จะทำการอัพเดตบิตในเว็บเพจ y ดังรูปแบบต่อไปนี้ $V_y \leftarrow V_x \; OR \; V_y$ ซึ่งในภาพที่ 19 แสดงถึงแสดงการคำนวณแบบวนรอบ

หลังจากคำนวณแบบวนรอบ *d* รอบแล้วนั้นบิตเวกเตอร์ของแต่ละเว็บเพจในเว็บกราฟจะแสดงเว็บเพจของแต่ละรอบในระยะทางน้อยกว่าหรือเท่ากับ *d* ซึ่งในงานวิจัยกล่าวว่าบิตเวกเตอร์ของเว็บเพจใดปรากฏ 1 เป็นจำนวนมากก็คาดได้ว่าเว็บเพจนั้นประกอบด้วยเว็บเพจช่วยเหลือในระยะทาง *d* เป็นจำนวนมาก หลังจากขั้นตอนนี้จะได้บิตเวกเตอร์ของแต่ละเว็บเพจในเว็บกราฟ โดยจะใช้มาตรวัดนี้ในการตัดสินเลือกว่าเว็บเพจใดที่ใช้วิธีการคำนวณด้วยวิธีเพจแร็งค์แบบตัดปลายต่อไป

2. เพจแร็งค์แบบตัดปลาย

วิธีการคำนวณเพจแร็งค์แบบตัดปลายมีแนวคิดพื้นฐานจากฟังค์ชั่นการจัดลำดับ (Baeza-Yates *et al*., 2006) ในการคำนวณหาค่าความสำคัญของแต่ละเว็บเพจที่แสดง โดยที่จะพิจารณาเว็บเพจบนอินเทอร์เน็ตเป็นเว็บกราฟที่มีทิศทาง $G = (\mathcal{V}, \mathcal{E})$ โดยกำหนดให้ $\mathcal{V}$ คือเซตของเว็บเพจในเว็บกราฟ G และ $\mathcal{E}$ คือเซตของลิงค์ที่เชื่อมต่อระหว่างเว็บเพจ เราจะสามารถเขียนเมตริกซ์

ความสัมพันธ์ P เป็นตัวแทนโครงสร้างความสัมพันธ์ของเว็บเพจต่างๆเหล่านั้นของเว็บกราฟได้ดังสมการต่อไปนี้

$$P(p,q) = \begin{cases} 1/\omega(q) & if \ (p,q) \in \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

กำหนดให้ค่าความสำคัญของแต่ละเว็บเพจที่แสดงโดยสมาชิกในเวกเตอร์ $\vec{W}$ และ $damping(t)$ คือฟังค์ชั่นลดค่า (decreasing function) จะสามารถแสดงฟังค์ชั่นการจัดลำดับดังรูปแบบสมการต่อไปนี้

$$\vec{W} = \sum_{t=0}^{\infty} \frac{damping(t)}{N} P^t \tag{20}$$

เมื่อพิจารณาการคำนวณฟังค์ชั่นการจัดลำดับที่กล่าวมาในสมการที่ (17) และการคำนวณเพจแร็งค์แล้วนั้นจะสามารถบอกได้ว่าเพจแร็งค์คือฟังค์ชั่นการจัดลำดับที่มีฟังค์ชั่นลดค่ารูปแบบเอ็กโพเน็นเชี่ยล (exponential decreasing) ดังสมการต่อไปนี้

$$damping(t) = (1-c) \times c^t \tag{21}$$

ความแตกต่างระหว่างการคำนวณเพจแร็งค์และการคำนวณเพจแร็งค์แบบตัดปลายนั้นคือเพจแร็งค์แบบตัดปลายจะไม่พิจารณาผลคะแนนเพจแร็งค์จากเว็บเพจในระยะทาง T โดยกำหนดให้ $T = d$ นำเสนอฟังค์ชั่นลดค่า $damping(t)$ ดังสมการต่อไปนี้

$$damping(t) = \begin{cases} 0 & t \le T \\ C\alpha^t & t > T \end{cases} \tag{22}$$

เมื่อ C คือค่าคงที่ใช้ในการทำนอมอลไลท์เพื่อให้เป็นไปตามสมการต่อไปนี้

$$\sum_{t=0}^{\infty} damping(t) = 1 \tag{23}$$

โดยในงานวิจัย URPC นั้นกำหนดให้ C มีค่าเท่ากับ $\frac{1-c}{c^{T+1}}$ ซึ่งส่งผลทำให้การคำนวณเพจแร็งค์แบบตัดปลายนั้นจะสามารถปรับลดผลกระทบของเว็บเพจที่ค่าคะแนนเพจแร็งค์ส่วนใหญ่ได้รับจากเว็บเพจในระยะทาง T

เมื่อสังเกตการณ์ทำงานของ URPC เป็นวิธีการค้นหาเว็บเพจที่มีเว็บเพจช่วยเหลือในระยะทาง T จำนวนมากโดยมีสมมติฐานว่าเว็บเพจลักษณะนี้เป็นการสแปมโครงสร้างลิงค์ และใช้การคำนวณเพจแร็งค์แบบตัดปลายเพื่อปรับลดค่าคะแนนเพจแร็งค์ของเว็บเพจเหล่านั้น เช่นเดียวกับงานวิจัยฉบับนี้ซึ่งพยายามปรับลดผลกระทบของการทำสแปมโครงสร้างลิงค์โดยการวิเคราะห์ลิงค์ แต่อย่างไรก็ตามในงานวิจัย URPC ยังมีจุดอ่อนคือถ้าทำการรวบรวมลิงค์โดยใช้วิธีการสร้าง Honey pot เนื่องจากลิงค์ของเว็บเพจเหล่านั้นจะได้จากเว็บเพจภายนอกสแปมฟาร์มซึ่งจะทำให้โครงสร้างของสแปมฟาร์มไม่เป็นไปตามสมมติฐาน และเมื่อสังเกตการณ์คำนวณเพจแร็งค์แบบตัดปลายแล้วนั้นเมื่อคำนวณกระทั่งรอบคำนวณ $i = \infty$ จะทำให้ค่าคะแนนเพจแร็งค์มีค่าเท่ากับ 0 เนื่องจากตัดผลคะแนนเพจแร็งค์จากเว็บเพจในระยะ T ดังนั้นในการคำนวณนั้นจำเป็นต้องมีทำการนอมอลไลท์เสมอ

อุปกรณ์และวิธีการ

อุปกรณ์

1. เครื่องคอมพิวเตอร์ 1 เครื่อง ประกอบด้วยอุปกรณ์ดังต่อไปนี้
   1.1. ซีพียู (CPU) เอเอ็มดี64 ความเร็ว 3 GHz
   1.2. หน่วยความจำหลัก 1 GB
   1.3. ฮาร์ดดิสต์ขนาด 80 GB
   1.4. การ์ดแลน

2. ซอฟแวร์ซึ่งประกอบด้วยโปรแกรมต่อไปนี้
   1.1. ระบบปฏิบัติการ Windows XP Professional
   1.2. ระบบปฏิบัติการ Linux
   1.3. Editor JAVA
   1.4. JDK 5.0
   1.5. Putty (SSH และ FTP)

วิธีการ

ภาพรวมของระบบ

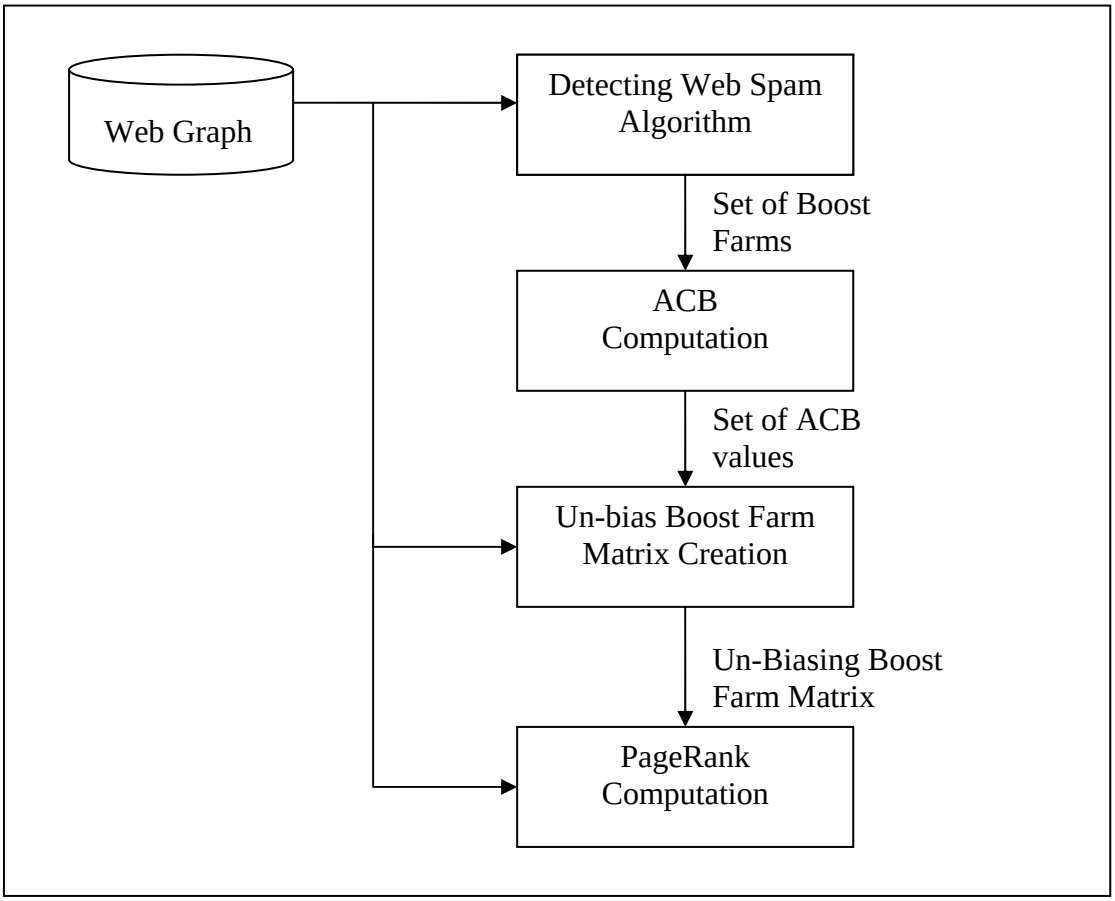
ดังที่กล่าวมาแล้วในส่วนของความรู้พื้นฐาน วิธีการสแปมโครงสร้างลิงค์นั้นสามารถกระทำได้หลายรูปแบบอีกทั้งบางโครงสร้างลิงค์ที่ดูคล้ายกับการสแปมนั้นเกิดจากการที่ผู้พัฒนาเว็บเพจที่มีความสนใจในหัวข้อลักษณะเดียวกันหรือลักษณะโครงสร้างทางธุรกิจสร้างลิงค์เชื่อมต่อซ้ำกันและกัน ดังนั้นในขั้นตอนแรกเราจะค้นหากลุ่มของเว็บเพจที่มีโครงสร้างลิงค์ที่คาดว่าเกิดจากการทำสแปมโครงสร้างลิงค์เว็บกราฟที่เราสนใจ โดยใช้วิธีการวิเคราะห์ลิงค์ของ (Wu and Davidson, 2005) เนื่องจากเราพบว่าเป็นวิธีการค้นหากลุ่มของเว็บเพจดังกล่าว อย่างอัตโนมัติ และเป็นวิธีการสร้างฐานข้อมูลเว็บกราฟด้วยวิธีคล้ายคลึงกันกับงานวิจัยที่เรากำลังสนใจอยู่ ดังนั้นเรา

จึงเห็นว่าการประยุกต์ใช้วิธีดังกล่าวในงานวิจัยของเรานั้นจึงเป็นวิธีการที่เหมาะสมกับการค้นหาบูซฟาร์ม เมื่อเราได้กลุ่มของบูซฟาร์มมาแล้ว เราจะทำการคำนวณหาอัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์ม (Average Change rate of probability of Boost farm: *ACB*) ของแต่ละบูซฟาร์ม โดยพิจารณาลักษณะโครงสร้างลิงค์ของบูซฟาร์ม และลิงค์ที่ชี้จากบูซฟาร์มไปยังเว็บเพจที่ไม่เป็นส่วนหนึ่งของบูซฟาร์ม เมื่อได้ค่าอัตราส่วนดังกล่าวมาแล้ว เราจะนำอัตราส่วนนี้มาทำการปรับลดผลกระทบของบูซฟาร์มโดยการสร้างเมตริกซ์ความสัมพันธ์ใหม่ โดยพิจารณาเปลี่ยนแปลงเฉพาะส่วนของเว็บเพจที่เกี่ยวข้องกับบูซฟาร์ม ก่อนที่จะนำเอาเมตริกซ์ดังกล่าวไปใช้สำหรับคำนวณหาค่าเพจแร็งค์ของเว็บเพจทั้งหมดในขั้นตอนต่อไป

นิยามที่ 1 บูซฟาร์ม (boost farm)

กำหนดให้บูซฟาร์มคือเครือข่ายของเว็บเพจที่มีลิงค์เชื่อมต่อกันอย่างหนาแน่น และเมื่อผู้ใช้งานที่เดินตามลิงค์ด้วยความน่าจะเป็นเท่าๆกันเสมอหลงท่องเข้ามายังบูซฟาร์มนั้นๆ จะมีความน่าจะเป็นน้อยที่ผู้ใช้งานนั้นๆจะสามารถเดินทางออกจากเครือข่ายของบูซฟาร์มนั้นได้ ซึ่งจะยังผลให้การคำนวณค่าคะแนนเพจแร็งค์ในบูซฟาร์มเหล่านั้นได้ค่าคะแนนสูงกว่าปกติ โดยปกติแล้วภายในเว็บกราฟใดๆอาจจะมีหรือไม่มีบูซฟาร์มอยู่ก็ได้ และถ้าเว็บกราฟนั้นมีบูซฟาร์มก็อาจมีได้มากกว่า 1 กลุ่ม

ก่อนที่จะทำความเข้าใจกับรายละเอียดนั้นสามารถที่จะดูภาพรวมของระบบได้จากภาพที่ 20 ซึ่งแสดงถึงขั้นตอนการทำงานตามลำดับของวิธีการที่นำเสนอในงานวิจัยฉบับนี้



ภาพที่ 20 ขั้นตอนของวิธีการปรับลดผลกระทบของบูซฟาร์มในการคำนวณเพจแร็งค์

วิธีการค้นหาการสแปมโครงสร้างลิงค์ (detecting web spam algorithm)

ในขั้นตอนนี้เราจะค้นหาการสแปมโครงสร้างลิงค์โดยประยุคใช้วิธีการของ (Wu and Davidson, 2005) โดยในงานวิจัยนี้จะใช้เฉพาะ 2 ขั้นตอนแรกคือ ขั้นตอนที่ 1 คือขั้นตอนการค้นหาคอมมอนเซต และขั้นตอนที่ 2 ขั้นตอนการขยายเซตเริ่มต้น ในการค้นหาเซตเริ่มต้นและเซตของลิงค์ฟาร์มโดยกำหนดค่าน้อยที่สุดที่ยอมรับได้ T_{IO}=3 และ T_{PP}=3 ตามผลการทดลองของ (Wu and Davidson, 2005) เมื่อได้เซตของเว็บเพจที่ถูกกำหนดจากวิธีการดังกล่าวว่าเป็นเซตของเว็บเพจที่ทำการสแปมโครงสร้างลิงค์ เราจะจัดกลุ่มเว็บเพจเหล่านั้นเป็นกลุ่มๆ โดยแต่ละกลุ่มแสดงถึงบูซฟาร์มใดๆ โดยใช้วิธีการคือ ถ้าเว็บเพจใดๆนั้นที่ถูกกำหนดว่าทำการสแปมโครงสร้างลิงค์มีลิงค์ชี้ไปยัง

เว็บเพจที่ถูกกำหนดว่าทำการสแปมโครงสร้างลิงค์เราจะรวมเว็บเพจทั้งสองนั้นอยู่ในกลุ่มเดียวกัน และถ้าเว็บเพจใดๆนั้นที่ถูกกำหนดว่าทำการสแปมโครงสร้างลิงค์มีลิงค์ถูกชี้โดยเว็บเพจที่ถูกกำหนดว่าทำการสแปมโครงสร้างลิงค์เราจะรวมเว็บเพจทั้งสองนั้นอยู่ในกลุ่มเดียวกันเช่นเดียวกับวิธีการแรกโดยในขั้นตอนเหล่านี้จะทำจนกว่าเซตของเว็บเพจที่ถูกกำหนดว่าเป็นเว็บเพจที่ทำการสแปมโครงสร้างลิงค์ได้ถูกจัดกลุ่มครบทุกเว็บเพจ

วิธีการคำนวณอัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์ม (ACB Computation)

จากสมมติฐานงานวิจัยฉบับนี้ที่กล่าวว่าเว็บเพจที่มีคุณภาพนั้นย่อมต้องมีลิงค์ชี้มาจากเว็บเพจที่มีคุณภาพจำนวนมากเช่นกัน ยังผลให้ในการทำสแปมโครงสร้างลิงค์ของอัลกอลิทึมเพจแร็งค์ส่วนใหญ่สามารถทำได้โดยวิธีการเก็บรวบรวมลิงค์ที่ชี้มาจากเว็บเพจที่ไม่มีคุณภาพจำนวนมากเพื่อทดแทนลิงค์ที่ชี้มาจากเว็บเพจที่มีคุณภาพ และทำการสะสมค่าคะแนนเพจแร็งค์ของเว็บเพจคุณภาพต่ำเหล่านั้นในรอบการคำนวณเพจแร็งค์ โดยสร้างโครงสร้างลิงค์ให้ส่งค่าคะแนนเพจแร็งค์ไปมาระหว่างเว็บเพจในกลุ่ม และในงานวิจัยเพจแร็งค์ (Page *et al*., 1998) ที่ได้กล่าวว่าลิงค์ที่ชี้มายังเว็บเพจที่พิจารณาสามารถแสดงถึงความสำคัญของเว็บเพจเหล่านั้น ดังนั้นในการคำนวณค่าอัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์ม เราจะพิจารณาเฉพาะความน่าจะเป็นที่บูซฟาร์มนั้นๆพยายามส่งค่ามอบค่าคะแนนเพจแร็งค์ให้แก่เว็บเพจอื่น โดยที่เราจะไม่พิจารณาปรับลดลิงค์ที่ชี้มายังกลุ่มของบูซฟาร์มนั้นๆ เพราะลิงค์ที่ชี้มายังกลุ่มของบูซฟาร์มเหล่านั้นจะเป็นตัวชี้วัดถึงคุณภาพของบูซฟาร์ม

ในกรณีที่สแปมฟาร์มประกอบด้วยลิงค์ชี้มาจากเว็บเพจที่มีคุณภาพจำนวนมาก จากสมมติฐานที่กล่าวมาแล้วนั้นทำให้วิธีการที่นำเสนอไม่สามารถแก้ปัญหานี้ได้ แต่ในความเป็นจริงแล้วนั้นถ้าผู้ที่ทำการสแปมโครงสร้างลิงค์สามารถรวบรวมหาลิงค์ชี้มาจากเว็บเพจที่มีคุณภาพจำนวนมากเพื่อเพิ่มค่าคะแนนเพจแร็งค์แก่เว็บเพจเป้าหมาย ก็จะไม่มีความจำเป็นในการสร้างโครงสร้างลิงค์เพื่อเพิ่มค่าคะแนนเพจแร็งค์ในรอบการคำนวณเพจแร็งค์แต่อย่างใด

ขั้นตอนการคำนวณอัตราการเปลี่ยนแปลงของความน่าจะเป็นเฉลี่ยของบูซฟาร์มหลังจากได้กลุ่มของบูซฟาร์มมาแล้ว เราจะสามารถแบ่งเว็บเพจในเว็บกราฟออกเป็น 2 กลุ่มใหญ่ๆคือ

เว็บเพจที่เป็นบูซฟาร์ม ซึ่งยังสามารถแยกย่อยได้อีก 2 ลักษณะคือ

1. เว็บเพจที่เป็นบูซฟาร์มที่มีลิงค์ชี้ไปยังเว็บเพจที่ไม่เป็นบูซฟาร์ม จากเว็บกราฟเริ่มต้นในภาพที่ 21 คือเว็บเพจหมายเลข 3 และ 6

2. เว็บเพจที่เป็นบูซฟาร์มที่ไม่มีลิงค์ชี้ไปยังเว็บเพจที่ไม่เป็นบูซฟาร์ม จากเว็บกราฟเริ่มต้นในภาพที่ 21 คือเว็บเพจหมายเลข 5

เว็บเพจที่ไม่เป็นบูซฟาร์ม ซึ่งยังสามารถแยกย่อยได้อีก 2 ลักษณะคือ

1. เว็บเพจที่ไม่เป็นบูซฟาร์มที่มีลิงค์ชี้ไปยังเว็บเพจที่เป็นบูซฟาร์ม จากเว็บกราฟเริ่มต้นในภาพที่ 21 คือเว็บเพจหมายเลข 1

2. เว็บเพจที่ไม่เป็นบูซฟาร์มที่ไม่มีลิงค์ชี้ไปยังเว็บเพจที่เป็นบูซฟาร์ม จากเว็บกราฟเริ่มต้นในภาพที่ 21 คือเว็บเพจหมายเลข 2, 4 และ 7

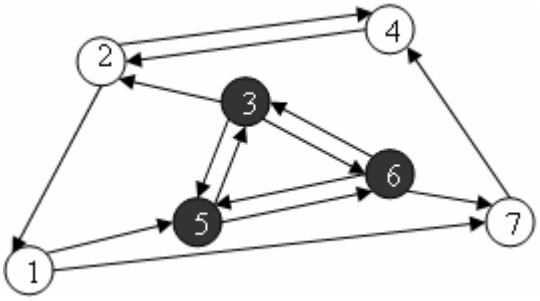
หลังจากที่เราแบ่งประเภทเว็บเพจแล้ว เราจะจำลองโครงสร้างเว็บกราฟจำลองเฉพาะกลุ่มของบูซฟาร์มใดๆ โดยมี 3 ขั้นตอนเพื่อแบ่งแยกโครงสร้างลิงค์ของบูซฟาร์มออกจากเว็บกราฟหลักและจำลองเว็บเพจเสมือนขึ้นมาเพื่อสะสมค่าคะแนนเพจแร็งค์ที่บูซฟาร์มส่งออกจากกลุ่มของตนเองโดยแบ่งเป็น 3 ขั้นตอนดังนี้คือ

1. ตัดแยกเว็บเพจที่เป็นบูซฟาร์มออกจากเว็บกราฟเริ่มต้น รวมถึงโครงสร้างลิงค์ที่เชื่อมต่อภายในกลุ่มของเว็บเพจที่เป็นบูซฟาร์ม

2. สร้างเว็บเพจเสมือน x ขึ้นมาโดยเว็บเพจเสมือน x จะมีลิงค์ชี้เข้าหาตนเอง เพื่อทำหน้าที่เป็นแร็งค์ซิงค์เก็บค่าคะแนนการคำนวณเพจแร็งค์ของเว็บเพจในกลุ่มบูซฟาร์มนั้นๆ

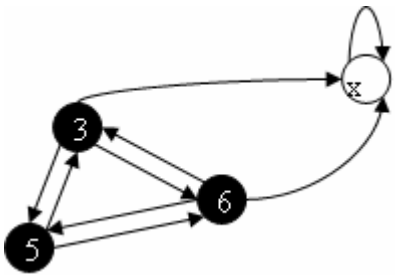
3. ทุกๆเว็บเพจที่เป็นบูซฟาร์มที่มีลิงค์ชี้ไปยังเว็บเพจที่ไม่เป็นบูซฟาร์มจะมีลิงค์ชี้มายังเว็บเพจเสมือน x จำนวนเท่ากับจำนวนลิงค์ที่ชี้ไปยังเว็บเพจที่ไม่ใช่บูซฟาร์ม

ซึ่งการจำลองโครงสร้างเว็บกราฟจำลองนี้ จะทำกับทุกๆกลุ่มของบูซฟาร์ม เช่นถ้าหากเรามีบูซฟาร์ม 3 กลุ่มในเว็บกราฟ เราจะต้องจำลองโครงสร้างเว็บกราฟจำลองทั้ง 3 เว็บกราฟด้วย เพื่อความเข้าใจยิ่งขึ้น เราขอยกตัวอย่างเว็บกราฟเริ่มต้นที่ประกอบด้วยบูซฟาร์ม 1 กลุ่ม ดังภาพที่ 21 โดยวงกลมสีขาวแสดงถึงเว็บเพจที่ไม่เป็นบูซฟาร์ม ส่วนวงกลมลงสีทึบแสดงถึงเว็บเพจที่ถูกพิจารณาว่าเป็นส่วนหนึ่งของบูซฟาร์ม



ภาพที่ 21 เว็บกราฟเริ่มต้น

จากเว็บกราฟเริ่มต้นภาพที่ 21 เราจะสามารถสร้างเว็บกราฟจำลองด้วยวิธีการที่กล่าวมาได้ดังภาพที่ 22



ภาพที่ 22 เว็บกราฟจำลอง

หลังจากเราสร้างโครงสร้างเว็บกราฟจำลองภาพที่ 22 แล้ว เราจะทำการกำหนดค่าเพจแร็งค์เริ่มต้นให้กับเว็บกราฟจำลองที่เราสร้างขึ้นใหม่นี้โดยให้ทุกๆเว็บเพจมีค่าเพจแร็งค์เริ่มต้นเท่ากับ $1/N$ เมื่อ N คือจำนวนเว็บเพจทั้งหมดในเว็บกราฟจำลอง หลังจากนั้นเราจะนำเว็บกราฟใหม่ที่ได้มาคำนวณค่าเพจแร็งค์เพื่อหาอัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์ม i นั้น (ACB_i)

นิยามที่ 2 อัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์ม (ACB)

อัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์ม i ใดๆ เขียนแทนด้วยสัญลักษณ์ ACB_i คือค่าที่แสดงถึง เมื่อมีผู้ใช้งานที่เดินตามลิงค์ด้วยความน่าจะเป็นเท่าๆกันเสมอ เดินทางมายังกลุ่มบูซฟาร์มกลุ่มที่ i ใดๆ จะมีอัตราส่วนของความน่าจะเป็นโดยเฉลี่ยที่จะเดินทางออกจากกลุ่มบูซฟาร์มกลุ่มนั้นไปยังเว็บเพจอื่นๆ มากน้อยเท่าใด จากนิยามที่กล่าวมาเราสามารถแสดงได้ดังสมการดังต่อไปนี้

$$ACB_i = \frac{1}{K_i}\sum_{n=1}^{K_i}(\Pr(i_{n-1}) - \Pr(i_n) + Randomjump(i_n)) / \Pr(i_{n-1}) \qquad (24)$$

โดยกำหนดให้เซตของเว็บเพจในบูซฟาร์มที่ i เขียนแทนด้วย BF_i จำนวนรอบการคำนวณเพจแร็งค์ของเว็บกราฟจำลองของบูซฟาร์มที่ i ใดๆเขียนแทนด้วย K_i ผลรวมของความน่าจะเป็นของทุกๆเว็บเพจในเว็บกราฟจำลองของบูซฟาร์มที่ i ใดๆยกเว้นเว็บเพจ x ในรอบการคำนวณเพจแร็งค์รอบที่ n เขียนแทนด้วย $Pr(i_n)$ และให้ค่าสุ่มกระโดดจากเว็บเพจ x ไปยังเว็บเพจใดๆนอกจากเว็บเพจ x ของเว็บกราฟจำลองของบูซฟาร์มที่ i ใดๆในรอบการคำนวณเพจแร็งค์รอบที่ n เขียนแทนด้วย $Randomjump(i_n)$

จากสมการที่ (24) เมื่อเรานิยามให้ $\vec{S}$ และ $\vec{D}$ คือแร็งค์เวกเตอร์ที่เป็นตัวแทนเวกเตอร์ $\vec{R}_i$ และ $\vec{R}_{i+1}$ ในสมการที่ (7) ตามลำดับ และ M คือเมตริกซ์ความสัมพันธ์ที่สร้างจากเว็บกราฟจำลองของบูซฟาร์มที่ i ใดๆ เราจะสามารถเขียนรหัสเทียม (psuedo code) ในการคำนวณค่า ACB_i ได้ดังภาพที่ 23

```
1 :   set S⃗ = [1/N]_{N×1}
2 :   set D⃗ = [0]_{N×1}
3 :   Pr(i_0) = Σ_{j=1, j≠row of webpage x}^{N} S⃗[j]
4 :   n = 0, sum = 0
5 :   while (error > δ){
6 :     D⃗ = 0.85(M × S⃗) + 0.15[1/N]_{N×1}
7 :     n = n + 1
8 :     Pr(i_n) = Σ_{j=1, j≠row of webpage x}^{N} D⃗[j]
9 :   Randomjump (i_n) = ((N − 1)/N)(0.15 × S⃗[j = row of webpage x])
10 :    sum = sum + (Pr(i_{n−1}) − Pr(i_n) + Randomjump (i_n)) / Pr(i_{n−1})
11 :    error = ||S⃗ − D⃗|| / ||S⃗||
12 :    S⃗ ← D⃗
13 :  }
14 :  ACB_i = sum / n
```

ภาพที่ 23 รหัสเทียมแสดงการคำนวณค่า ACB i ของเว็บเพจในบูซฟาร์มที่ i ใดๆ

ค่าของ ACB_i ที่คำนวณได้นั้นจะมีค่าอยู่ในขอบเขต [0,1] บทพิสูจน์โดยละอียดได้นำเสนอไว้ในภาคผนวก ก เนื่องจากถ้าเราพิจารณาจากค่า ACB_i ใดๆที่คำนวณได้จากสมการที่ (24) นั้นจะเป็นค่าเฉลี่ยของอัตราส่วน

$$\frac{\Pr(i_{n-1}) - (\Pr(i_n) - Ramndomjump(i_n))}{\Pr(i_{n-1})}$$

ซึ่งค่าอัตราส่วนนี้จะมีค่าเป็นบวกเสมอเนื่องจากวิธีการสร้างโครงสร้างกราฟจำลองจะมีเว็บเพจเสมือน x ที่ทำหน้าที่เป็นแร็งค์ซิงค์ทำให้ค่า $\Pr(i_n) - Ramndomjump(i_n)$ ย่อมมีค่าน้อยกว่าหรือเท่ากับค่า $\Pr(i_n)$ เสมอ

ในกรณีค่า ACB_i ใดๆนั้นมีค่าเท่ากับ 0 แสดงว่าเว็บเพจในกลุ่มของบูซฟาร์มที่ i ใดๆ มีค่าผลรวมความน่าจะเป็น $\Pr(i_n)$ และ $\Pr(i_n) - Ramndomjump(i_n)$ มีค่าเท่ากัน หรืออาจกล่าวได้ว่าบูซฟาร์มกลุ่มที่ i ใดๆนั้นมีลักษณะเป็นแร็งค์ซิงค์เมื่อผู้ใช้งานที่เดินตามลิงค์ด้วยความน่าจะเป็นเท่าๆกันเสมอเดินทางมายังกลุ่มบูซฟาร์มกลุ่มที่ i ใดๆนั้นแล้วจะมีความน่าจะเป็นเฉลี่ยเป็น 0 ที่สามารถจะเดินทางออกจากกลุ่มของบูซฟาร์มนั้นได้ (หรือติดอยู่ในบูซฟาร์มนั่นเอง) ซึ่งในความเป็นจริงแล้วนั้นจะเกิดเหตุการณ์นี้ขึ้นได้ยากเนื่องจากในอัลกอลิทึมเพจแร็งค์เมื่อพิจารณาเว็บเพจใดแล้วจะมีการสุ่มกระโดดด้วยความน่าจะเป็น $0.15/N^*$ (เมื่อ N^* คือจำนวนเว็บเพจทั้งหมดในเว็บกราฟ) ไปยังทุกๆเว็บเพจในเว็บกราฟทำให้ ACB_i จะไม่มีค่าเท่ากับ 0 ยกเว้นในกรณีที่เว็บกราฟที่พิจารณาเมื่อแสดงโดยเมตริกความสัมพันธ์แล้วนั้นมีคุณสมบัติเป็นเมตริกซ์เอกลักษณ์ หรือเมตริกที่ผลรวมในแต่ละแถวมีค่าเท่ากับ 1 ในกรณีกำหนดค่า $\vec{R}_0$ มีค่าเท่ากันในทุกๆสมาชิก

และในกรณีค่า ACB_i ใดๆนั้นมีค่าเท่ากับ 1 แสดงว่าเมื่อผู้ใช้งานที่เดินตามลิงค์ด้วยความน่าจะเป็นเท่าๆกันเสมอเดินทางมายังกลุ่มบูซฟาร์มกลุ่มที่ i ใดๆนั้นแล้วมีความน่าจะเป็นสูงมากที่จะเดินทางออกจากกลุ่มบูซฟาร์มกลุ่มที่ i ใดๆนั้นออกไปยังเว็บเพจอื่นๆในทันที ซึ่งในความเป็นจริงแล้วนั้นจะไม่เกิดเหตุการณ์นี้ขึ้นเนื่องจากในอัลกอลิทึมเพจแร็งค์เมื่อพิจารณาเว็บเพจใดแล้วจะการสุ่มกระโดดด้วยความน่าจะเป็น $0.15/N^*$ ไปยังทุกๆเว็บเพจในเว็บกราฟซึ่งรวมถึงเว็บเพจ ตนเองและเว็บเพจที่ถูกจัดอยู่ในกลุ่มของบูซฟาร์มทำให้ ACB_i จะไม่มีค่าเท่ากับ 1

การสร้างเมตริกซ์ความสัมพันธ์แบบใหม่ (Un-bias Boost Farm Matrix Creation)

วิธีการในขั้นตอนนี้จะทำต่อจากขั้นตอนการคำนวณหาอัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์มดังที่ได้กล่าวมาแล้ว ซึ่งภายในขั้นตอนนี้เราจะนิยามเมตริกซ์ขนาด N x N ขึ้นใหม่อีก 3 ชนิด (เมื่อ N คือจำนวนเว็บเพจทั้งหมดในเว็บกราฟ) คือ

1. เมตริกซ์บูซฟาร์ม (boost farm matrix)

2. เมตริกซ์ที่ไม่มีบูซฟาร์ม (non-boost farm matrix)

3. เมตริกซ์ลิงค์เสมือน (virtual link matrix)

นิยามที่ 3 เมตริกซ์บูซฟาร์ม (boost farm matrix)

เมตริกซ์บูซฟาร์มของบูซฟาร์มกลุ่มที่ i ใดๆ เขียนแทนด้วยสัญลักษณ์ BFM_i เป็นเมตริกซ์ที่เกิดขึ้นจากการแทนทุกๆ คอลัมน์ที่ไม่เป็นสมาชิกของบูซฟาร์มกลุ่มที่ i ใดๆ ในเมตริกซ์ความสัมพันธ์เดิม M เป็นค่าศูนย์ทั้งคอลัมน์ ซึ่งแสดงได้ดังสมการดังต่อไปนี้

$$BFM_i(p,q) = \begin{cases} 1/\omega(q) & if\ (q,p) \in \mathcal{E} \text{ and } q \in BF_i \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

จากส่วนของเว็บกราฟในภาพ ที่ 21 ที่แสดงถึงเว็บกราฟเริ่มต้นนั้น เราสามารถแสดงตัวอย่างของเมตริกซ์บูซฟาร์มจากสมการที่ (25) ได้ดังภาพที่ 24

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 1/3 & 0 \end{bmatrix}$$

ภาพที่ 24 เมตริกซ์บูซฟาร์มที่สร้างขึ้นจากเว็บกราฟเริ่มต้นในรูปที่ 21

นิยามที่ 4 เมตริกซ์ที่ไม่มีบูซฟาร์ม (non-boost farm matrix)

เมตริกซ์ที่ไม่มีบูซฟาร์มเขียนแทนด้วยสัญลักษณ์ *non_BFM* จะเป็นเมตริกซ์ที่เกิดขึ้นจากการแทนค่าทุกๆคอลัมน์ที่เป็นสมาชิกบูซฟาร์มทั้งหมดที่หาได้ในเว็บกราฟในเมตริกซ์ความสัมพันธ์เดิม M ด้วยค่าศูนย์ทั้งคอลัมน์ ซึ่งสามารถแสดงได้ดังสมการดังต่อไปนี้

$$\text{non\_}BFM(p,q) = \begin{cases} 1/\omega(q) & if\ (q,p) \in \mathcal{E} \text{ and } q \notin BF \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

จากส่วนของเว็บกราฟในภาพที่ 21 ที่แสดงถึงเว็บกราฟเริ่มต้นนั้น เราสามารถแสดงตัวอย่างเมตริกซ์ที่ไม่มีบูซฟาร์มจากสมการที่ (26) ได้ดังภาพที่ 25

$$\begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 & 1 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

ภาพที่ 25 เมตริกซ์ที่ไม่มีบูซฟาร์มที่สร้างขึ้นจากเว็บกราฟเริ่มต้นในรูปที่ 21

ถ้าเราพิจารณาสมการที่ (2), (6), (25) และ (26) ร่วมกัน จะเห็นได้ว่า เมตริกซ์ความสัมพันธ์เดิม M สามารถถูกแสดงในรูปของเมตริกซ์ใหม่ทั้งสองที่พึ่งกล่าวมาได้ดังสมการดังต่อไปนี้

$$M = \left[non\_BFM + \sum_{i=1}^{n} BFM_i \right] + (\vec{d} \times \vec{E}^T)^T \qquad (27)$$

นิยามที่ 5 ลิงค์เสมือน (virtual link)

ลิงค์เสมือนคือลิงค์ที่สร้างเพิ่มให้แก่กลุ่มของบูซฟาร์มใดๆเพื่อยังผลให้คุณสมบัติการเพิ่มค่าคะแนนเพจแร็งค์ของบูซฟาร์มกลุ่มนั้นๆลดลง ซึ่งในขั้นตอนการสร้างลิงค์เสมือนนั้นเราจะพิจารณาบูซฟาร์มทีละกลุ่ม โดยเสมือนว่าเราต่อลิงค์เสมือนไปยังทุกๆ เว็บเพจที่ไม่ได้เป็นสมาชิกของบูซฟาร์มกลุ่มนั้น ซึ่งในขั้นตอนการทำงานจริงนั้นเราจะไม่มีการเพิ่มลิงค์เสมือนเข้าไปในเว็บกราฟแต่อย่างใด แต่เราจะสร้างเมตริกลิงค์เสมือน ที่เป็นตัวแทนลิงค์เสมือนของบูซฟาร์มกลุ่มนั้น เสมือนว่ามีลิงค์เหล่านั้นอยู่ในเว็บกราฟจริง และใช้เมตริกลิงค์เสมือนปรับลดผลกระทบของบูซฟาร์มในการคำนวณเพจแร็งค์

นิยามที่ 6 เมตริกซ์ลิงค์เสมือน (virtual link matrix)

ส่วนเมตริกซ์ลิงค์เสมือนของเว็บเพจที่เป็นสมาชิกของบูซฟาร์มกลุ่มที่ i ใดๆ เขียนแทนด้วยสัญลักษณ์ VM_i คือเมตริกซ์ที่พิจารณาเฉพาะเว็บเพจที่เป็นบูซฟาร์ม โดยเสมือนว่าเราต่อ “ลิงค์เสมือน” ไปยังทุกๆ เว็บเพจที่ไม่ได้เป็นสมาชิกของบูซฟาร์มกลุ่มที่ i ใดๆ นั้น ด้วยค่าความน่าจะเป็นเท่ากับ $1/(N\text{-}|BF_i|)$ เมื่อ $(N\text{-}|BF_i|)$ คือจำนวนเว็บเพจทั้งหมดที่ไม่รวมเว็บเพจภายในกลุ่มของบูซฟาร์มที่ถูกพิจารณาอยู่ และแทนทั้งคอลัมน์ของคอลัมน์ที่ไม่ใช่เป็นสมาชิกของเซตของเว็บเพจในบูซฟาร์มกลุ่มที่ i ใดๆด้วยค่าศูนย์ซึ่งสามารถแสดงได้ดังสมการดังต่อไปนี้

$$VM_i(p,q)=\begin{cases}1/(N-|BF_i|) & if\ p\notin BF_i \text{ and} q\in BF_i\\ 0 & \text{otherwise}\end{cases} \qquad (28)$$

ในทำนองเดียวกัน จากส่วนของเว็บกราฟในภาพที่ 21 ที่แสดงถึงเว็บกราฟเริ่มต้นนั้น เราสามารถเขียนเมตริกซ์ลิงค์เสมือนจากสมการที่ (28) ได้ดังภาพที่ 26

$$\begin{bmatrix} 0 & 0 & 1/4 & 0 & 1/4 & 1/4 & 0\\ 0 & 0 & 1/4 & 0 & 1/4 & 1/4 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 1/4 & 0 & 1/4 & 1/4 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 1/4 & 0 & 1/4 & 1/4 & 0 \end{bmatrix}$$

ภาพที่ 26 เมตริกซ์ที่ลิงค์เสมือนที่สร้างขึ้นจากเว็บกราฟเริ่มต้นในรูปที่ 21

ในกรณีที่ $|BF_i|$ มีค่าเท่ากับ N จะทำให้ทุกๆสมาชิกในเมตริกซ์ลิงค์เสมือนมีค่าเท่ากับ 0 ซึ่งจะทำให้วิธีการที่นำเสนอมาทำงานผิดพลาด นั่นคือเมื่อคำนวณสมการที่ (30) จนรอบการคำนวณ $i=\infty$ จะทำให้ทุกๆสมาชิกของแร็งค์เวกเตอร์จะมีค่าลู่เข้าสู่ 0 แต่อย่างไรก็ตามเหตุการณ์ในกรณีนี้จะเกิดขึ้นก็ต่อเมื่อเว็บเพจทั้งหมดในอินเทอร์เน็ตเป็นบูซฟาร์มทั้งหมด ซึ่งในความเป็นจริงนั้นเป็นเรื่องยากที่จะเกิดเหตุการณ์นี้ขึ้น

จากนิยามของเมตริกซ์ทั้งสามที่ได้กล่าวมาแล้วนั้น เมื่อกำหนดให้ n คือจำนวนบูซฟาร์มทั้งหมดในเว็บกราฟ เราจะสามารถแสดงวิธีการปรับลดผลกระทบของบูซฟาร์มโดยการสร้างลิงค์เสมือนผ่านการสร้างเมตริกซ์ความสัมพันธ์อันใหม่ M_b ได้ดังสมการต่อไปนี้

$$M_b = \left\{ non\_BFM + \sum_{i=1}^{n} [(1 - ACB_i)VM_i + (ACB_i)BFM \ ] \right\} + (\vec{d} \times \vec{E}^T)^T \qquad (29)$$

ดังนั้นจากตัวอย่างดังภาพที่ 24, 25 และ 26 ที่แสดงถึงเมตริกซ์บูซฟาร์ม เมตริกซ์ที่ไม่มีบูซฟาร์ม และเมตริกซ์ลิงค์เสมือนที่สร้างขึ้นจากส่วนของเว็บกราฟตัวอย่างในภาพที่ 21 นั้น เราสามารถแสดงตัวอย่างของเมตริกซ์ความสัมพันธ์ที่ปรับลดผลกระทบของบูซฟาร์ม M_b จากสมการที่ (29) ได้ดังภาพที่ 27

0	1/2	(1-ACB)/4	0	(1-ACB)/4	(1-ACB)/4	0
0	0	(1-ACB)/4+ACB/3	1	(1-ACB)/4	(1-ACB)/4	0
0	0	0	0	ACB/2	ACB/3	0
0	1/2	(1-ACB)/4	0	(1-ACB)/4	(1-ACB)/4	1
1/2	0	ACB/3	0	0	ACB/3	0
0	0	ACB/3	0	ACB/2	0	0
1/2	0	(1-ACB)/4	0	(1-ACB)/4	(1-ACB)/4+ACB/3	0

ภาพที่ 27 เมตริกซ์ความสัมพันธ์ที่ปรับลดผลกระทบของบูซฟาร์ม Mbที่สร้างขึ้นจากเว็บกราฟเริ่มต้นในรูปที่ 21

เมตริกซ์ความสัมพันธ์ใหม่ M_b ที่ผ่านการปรับลดผลกระทบของบูซฟาร์มโดยการสร้างลิงค์เสมือนจากเมตริกซ์ VM_i นั้นมีความแตกต่างจากเมตริกซ์ความสัมพันธ์เดิม M ตรงที่มีการเพิ่มลิงค์เสมือนพร้อมทั้งปรับลดผลกระทบที่เกิดจากอัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์มไว้ด้วย และด้วยโครงสร้างลิงค์ที่แทนด้วยเมตริกซ์ความสัมพันธ์ที่ปรับลดผลกระทบของบูซฟาร์ม M_b นี้ จะยังผลทำให้สามารถกระจายความน่าจะเป็นที่ผู้ใช้งานที่เดินตามลิงค์ด้วยความน่าจะเป็นเท่าๆกันเสมอจะท่องวนติดอยู่ภายในกลุ่มของบูซฟาร์มนั้นๆสามารถออกไปยังเว็บเพจภายนอกบูซฟาร์มอื่นๆได้ ทำให้คุณสมบัติของบูซฟาร์มที่จะเพิ่มค่าคะแนนเพจแร็งค์จากเว็บเพจคุณภาพต่ำจำนวนมากเสียไป อีกทั้งเมตริกซ์ความสัมพันธ์ที่ถูกสร้างขึ้นมาใหม่นั้นยังคล้อยตามคุณสมบัติมาคอร์ฟ ซึ่งเราสามารถนำไปใช้คำนวณค่าเพจแร็งค์ด้วยวิธีการดั้งเดิม และรับประกันได้

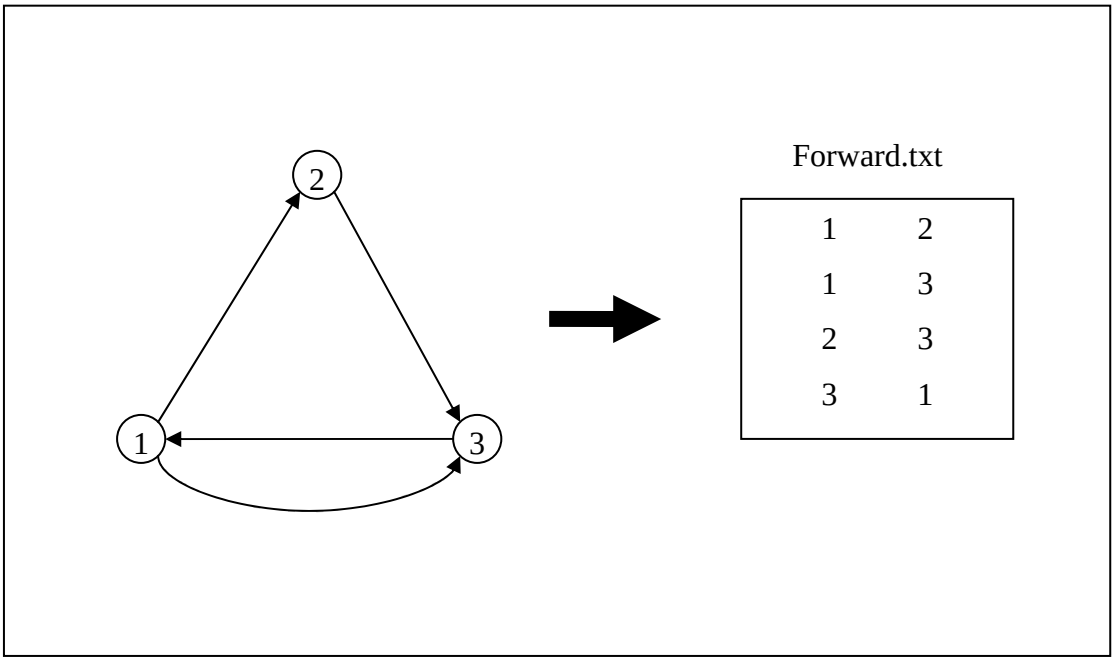
ว่าเมื่อจำนวนรอบในการคำนวณ $i = \infty$ แล้วได้คำตอบของแร็งค์เวกเตอร์ลู่เข้าสู่ค่าๆหนึ่งเสมอ ซึ่งสามารถเขียนได้ดังสมการต่อไปนี้

$$\vec{R}_{i+1} = cM_b \times \vec{R}_i + (1-c) \times [\frac{1}{N}]_{N\ x\ 1} \qquad (30)$$

ผลและวิจารณ์

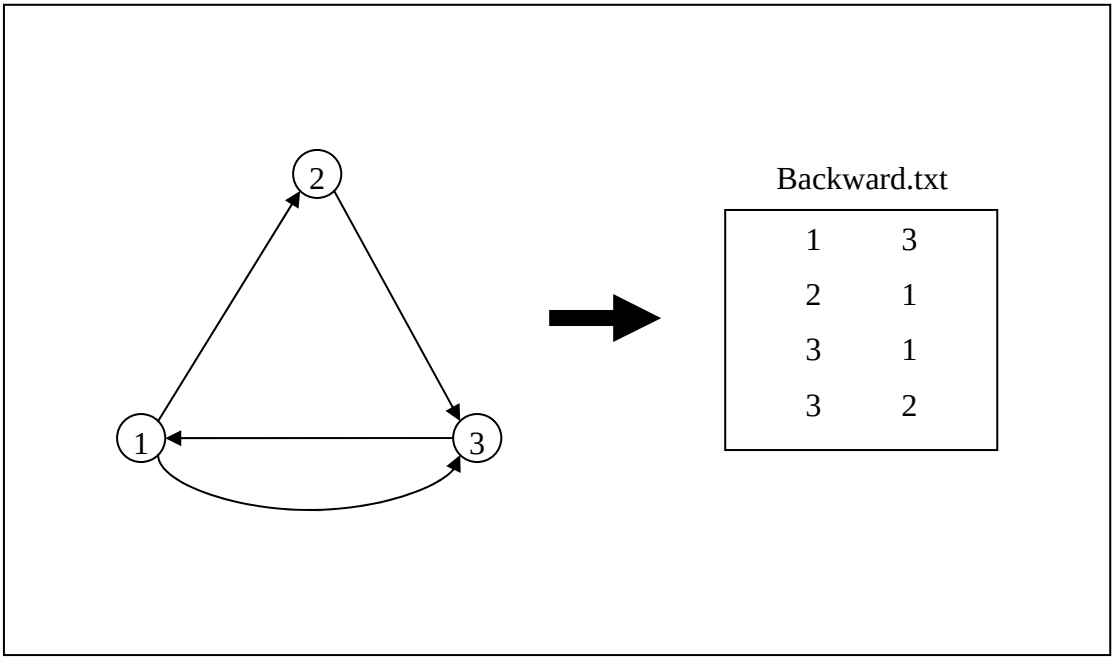
การเตรียมข้อมูล

เพื่อทดสอบวิธีการที่นำเสนอ เราทำการเก็บรวบรวมฐานข้อมูลเว็บกราฟมาจากระบบสืบค้นข้อมูลยาฮู (Yahoo, 2007) โดยใช้ยาฮูเอพีไอ (Yahoo API) และใช้คำสืบค้นรูปแบบต่างๆ ที่ถูกนำไปใช้ในการสร้างฐานข้อมูลในงานวิจัยของ (Wu and Davidson, 2005a) สืบค้นไปยังระบบสืบค้นข้อมูลยาฮูและนำผลค้นคืน 20 ลำดับแรกเป็นเซตของเว็บเพจเริ่มต้น หลังจากนั้นจึงทำขยายกลุ่มเว็บเพจเริ่มต้นโดยการสร้าง เบสเซตจากเซตของเว็บเพจเริ่มต้นโดยจากทุกเว็บเพจในเซตเริ่มต้นนั้น เราจะสืบค้นเว็บเพจ 50 ลำดับแรกที่มีลิงค์ชี้มายังเว็บเพจเริ่มต้นด้วยระบบสืบค้นข้อมูลยาฮูเช่นกัน และเมื่อได้เว็บกราฟจากคำสืบค้นใดๆแล้วนั้น จะทำการกำหนดหมายเลขประจำเว็บเพจแต่ละอันและแสดงโครงสร้างลิงค์ให้อยู่ในรูปแบบคู่ลำดับของหมายเลขประจำเว็บเพจดังตัวอย่างแสดงในภาพ 28



ภาพที่ 28 วิธีการสร้างลิงค์ไฟล์

เมื่อเราได้รูปแบบคู่ลำดับของหมายเลขประจำเว็บเพจแล้วนั้น เราจะได้โครงสร้างลิงค์ที่แสดงว่าเว็บเพจใดๆ นั้นมีลิงค์ชี้ไปยังเว็บเพจใด หลังจากนั้นเราจะสร้างคู่ลำดับที่แสดงว่าเว็บเพจใดๆนั้นถูกลิงค์มาจากเว็บเพจใดจากเว็บกราฟ ดังภาพที่ 29



ภาพที่ 29 วิธีการสร้างแบ็คลิงค์ไฟล์

เมื่อได้คู่ลำดับของหมายเลขประจำเว็บเพจที่แสดงว่าเว็บเพจใดๆมีลิงค์ชี้ไปยังเว็บเพจใด และคู่ลำดับที่แสดงว่าเว็บเพจใดๆถูกลิงค์มาจากเว็บเพจใดมาแล้วนั้น ขั้นตอนต่อไปเราจะเริ่มทำการทดลองปรับลดผลกระทบของบูซฟาร์มในการคำนวณเพจแร็งค์

ผล

การทดลองบนเว็บกราฟที่ค้นคืนจาก Yahoo API

ในการวัดผลการทดลองในขั้นตอนนี้เราจะสร้างฐานข้อมูลเว็บกราฟจากระบบสืบค้นข้อมูล Yahoo (Yahoo, 2007) โดยฐานข้อมูลที่สร้างนี้มีขนาด 250232 เว็บเพจ และใช้คำค้นหลักเป็น “credit card applications” เหมือนกับที่ใช้ในการทดลองในบทความที่ Wu และ Davidson

(2005a) ได้นำเสนอไว้ แล้วจึงคำนวณหาค่าคะแนนเพจแร็งค์ของเว็บกราฟที่เก็บรวบรวมได้ผลลัพธ์ดังตารางที่ 1

ตารางที่ 1 คะแนนเพจแร็งค์ 30 ลำดับแรกของเว็บกราฟก่อนทำการสแปมโครงสร้างลิงค์

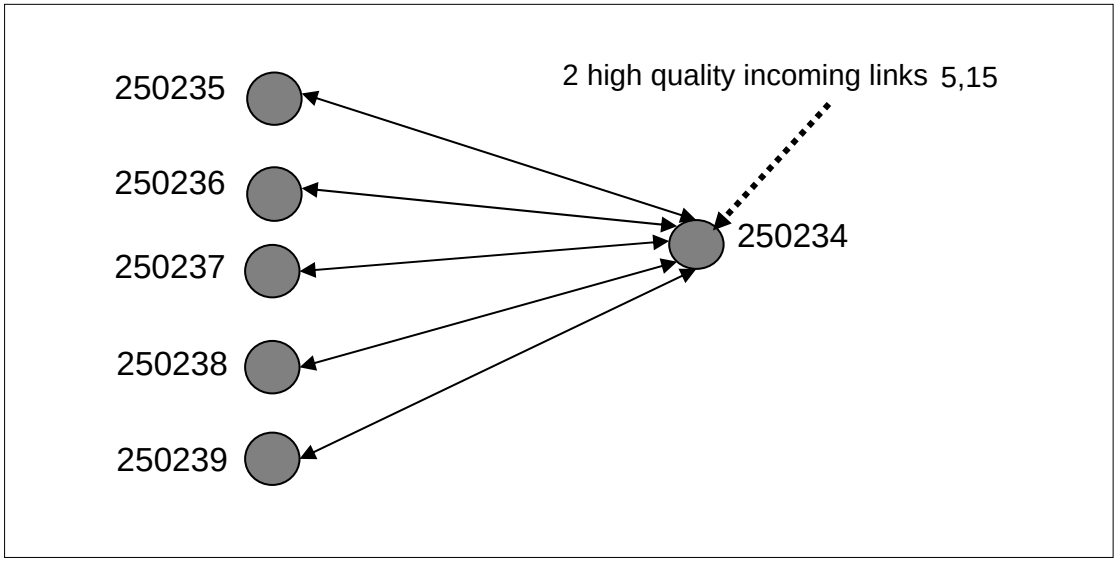
Rank	Page ID	Page Rank Score	Rank	Page ID	Page Rank Score
1	4	0.02083	16	9	0.00376
2	659	0.01635	17	343	0.00372
3	648	0.00885	18	6	0.00371
4	11	0.00882	19	3	0.00366
5	84	0.00818	20	499	0.00338
6	1838	0.00797	21	80	0.00302
7	1	0.00797	22	3041	0.00302
8	1829	0.00766	23	3042	0.00302
9	47	0.00766	24	22	0.00269
10	1830	0.00766	25	163	0.00260
11	2935	0.00535	26	652	0.00252
12	2934	0.00516	27	660	0.00251
13	2	0.00516	28	17	0.00238
14	5	0.00384	29	456	0.00236
15	15	0.00382	30	251	0.00233

หลังจากนั้นเราจะทำการสแปมโครงสร้างลิงค์โดยการเพิ่มสแปมฟาร์มซึ่งถูกนำเสนอโดย (Gyongyi and Garcia-Molina, 2005b) ว่าโครงสร้างนั้นๆเป็นการแสปมโครงสร้างลิงค์ที่มีผลกระทบต่อการคำนวณเพจแร็งค์จริง โดยเราจะเพิ่มจำนวน 3 รูปแบบซึ่งเป็นตัวแทนของการสแปมโครงสร้างลิงค์แบบ 1, *n* กลุ่ม และเว็บวงแหวน สาเหตุที่ทำการสแปมเว็บเพจเพิ่มเติมเนื่องจากในงานวิจัยฉบับนี้นั้นพิจารณาเฉพาะการปรับลดผลกระทบของบูซฟาร์มในการคำนวณเพจแร็งค์โดยที่ไม่รวมถึง

ขั้นตอนการค้นหาบูซฟาร์ม และเราต้องการให้เห็นถึงการเปลี่ยนแปลงของลำดับเว็บเพจในสแปมฟาร์มตัวอย่าง

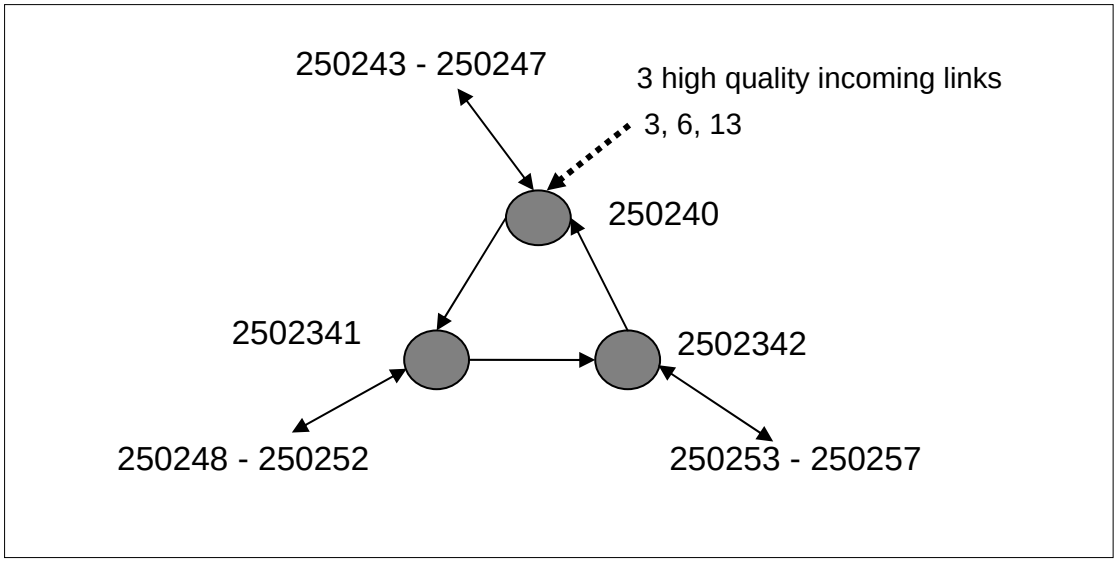
เว็บเพจที่เราทำการสแปมเพิ่มได้แก่เว็บเพจหมายเลข 250234 - 250262 โดยเว็บเพจเหล่านี้จะมีลิงค์จากเว็บเพจที่ไม่เป็นสแปมฟาร์มชี้เข้ามาแต่ละเว็บเพจ เว็บเพจละ 10 ลิงค์ และลิงค์ที่ชี้เข้ามานี้มาจากเว็บเพจที่มีค่าเพจแร็งค์ต่ำสุด 90 อันดับแรก แต่จะมีเว็บเพจ 3 เว็บเพจที่มีลิงค์ชี้มาจากเว็บเพจที่มีคุณภาพนั่นคือเว็บลเพจที่ 250234, 250240 และ 250260 แสดงได้ดังรูปต่อไปนี้

1. โครงสร้างสแปมฟาร์มที่ 1 ประกอบด้วยเว็บเพจ 250234 – 250239 โดยที่เว็บเพจ 250234 จะลิงค์ชี้เพิ่มจากเว็บเพจหมายเลขที่ 5 และ 15 ซึ่งเป็นเว็บเพจที่มีค่าเพจแร็งค์ลำดับที่ 14 และ 15 ตามลำดับดังภาพที่ 30



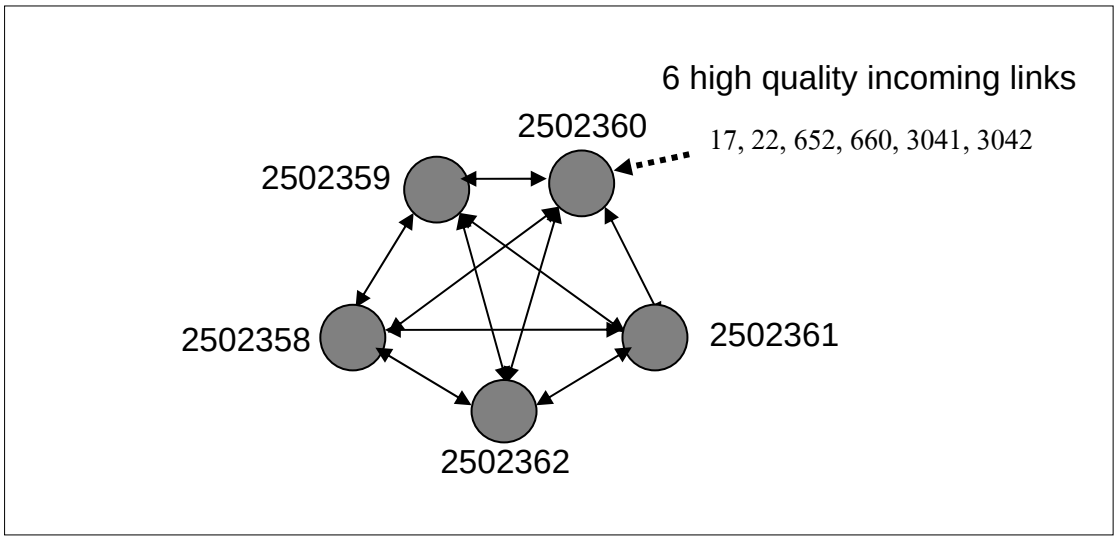
ภาพที่ 30 สแปมฟาร์มกลุ่มที่ 1 ที่เป็นตัวแทนการสแปมโครงสร้างลิงค์แบบ 1 กลุ่ม

2. โครงสร้างสแปมฟาร์มที่ 2 ประกอบด้วยเว็บเพจ 250240 – 250257 โดยที่เว็บเพจ 250240 จะลิงค์ชี้เพิ่มจากเว็บเพจหมายเลขที่ 3, 6 และ 343 ซึ่งเป็นเว็บเพจที่มีค่าเพจแร็งค์ลำดับที่ 19, 18 และ 17 ตามลำดับดังภาพที่ 31



ภาพที่ 31 สแปมฟาร์มกลุ่มที่ 2 ที่เป็นตัวแทนเว็บวงแหวน

3. โครงสร้างสแปมฟาร์มที่ 3 ประกอบด้วยเว็บเพจ 250258 – 250262 โดยที่เว็บเพจ 250260 จะลิงค์ชี้เพิ่มจากเว็บเพจหมายเลขที่ 17, 22, 652, 660, 3041 และ 3042 ซึ่งเป็นเว็บเพจที่มีค่าเพจแร็งค์ลำดับที่ 28, 24, 26, 27, 22 และ 23 ตามลำดับดังภาพที่ 32



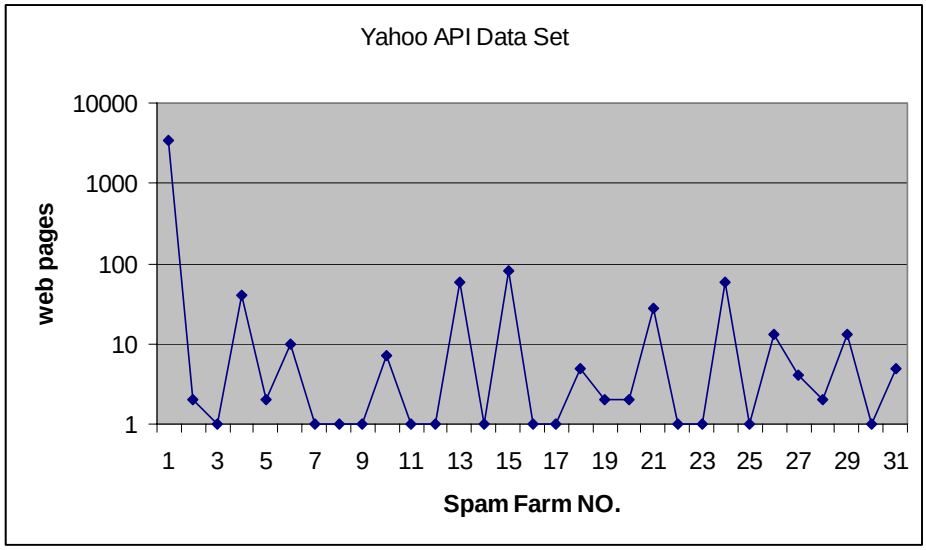
ภาพที่ 32 สแปมฟาร์มกลุ่มที่ 2 ที่เป็นตัวแทนการสแปมโครงสร้างลิงค์แบบ *n* กลุ่ม

หลังจากนั้นทำการคำนวณค่าเพจแร็งค์ให้กับเว็บเพจในฐานข้อมูลนี้มีค่าดังตารางที่ 2 โดยแถวที่มี (A) ในลำดับคือเว็บเพจที่ถูกเพิ่มเข้าสู่เว็บกราฟและกำหนดให้เป็นสแปมฟาร์ม และแถวที่มี (B) ในลำดับคือเว็บเพจที่ถูกค้นพบโดยวิธีการ (Wu and Brian, 2005a) ว่าเป็นสแปมฟาร์ม

ตารางที่ 2 คะแนนเพจแร็งค์ 30 ลำดับแรกของเว็บกราฟหลังทำการสแปมโครงสร้างลิงค์

Rank	Page ID	Page Rank Score	Rank	Page ID	Page Rank Score
1	4	0.01998	16 B	2	0.00437
2 A	250234	0.01272	17 A	250262	0.00410
3	659	0.01239	18 A	250259	0.00410
4 A	250240	0.00971	19 A	250261	0.00410
5 A	250260	0.00771	20 A	250258	0.00410
6	1838	0.00737	21	5	0.00367
7	1	0.00737	22 B	9	0.00360
8	1829	0.00710	23	3	0.00351
9 B	47	0.00710	24 A	250241	0.00346
10	1830	0.00710	25 B	499	0.00324
11	648	0.00671	26 B	163	0.00250
12	11	0.00668	27	22	0.00245
13 B	2935	0.00451	28	652	0.00241
14 B	84	0.00451	29	660	0.00241
15 B	2934	0.00437	30 B	80	0.00239

ซึ่งในขั้นตอนวิธีการ (Wu and Brian, 2005a) ได้พบกลุ่มเว็บเพจที่เป็นสแปมฟาร์ม 31 กลุ่มด้วยกัน โดยแต่ละกลุ่มมีจำนวนเว็บเพจดังภาพที่ 33



ภาพที่ 33 กราฟแสดงจำนวนเว็บเพจที่เป็นสมาชิกของสแปมฟาร์มแต่ละกลุ่ม

หลังจากนั้นได้นำวิธีการที่นำเสนอมาปรับลดอิทธิพลของบูซฟาร์ม แล้วไปคำนวณค่าเพจแร็งค์ตามสมการที่ (30) ส่งผลให้ค่าคะแนนเพจแร็งค์ของเว็บเพจ 30 ลำดับแรกจากตารางที่ 2 มีการเปลี่ยนแปลงค่าคะแนนเพจแร็งค์ดังตารางที่ 3 ต่อไปนี้

ตารางที่ 3 การเปลี่ยนแปลงคะแนนเพจแร็งค์ 30 ลำดับแรกจากตารางที่ 2 หลังจากปรับลดลดผลกระทบของบูซฟาร์มในการคำนวณเพจแร็งค์

Page ID	Rank	Un-bias Rank	Page Rank Score	Un-bias PageRank Score	Page ID	Rank	Un-bias Rank	Page Rank Score	Un-bias PageRank Score
4	1	1	0.01998	0.02624	2 B	16	15	0.00437	0.00425
250234 A	2	3	0.01272	0.01378	250262 A	17	8101	0.00410	0.00000510
659	3	2	0.01239	0.01651	250259 A	18	8102	0.00410	0.00000510
250240 A	4	12	0.00971	0.00464	250261 A	19	8129	0.00410	0.00000496
250260 A	5	17	0.00771	0.00357	250258 A	20	8130	0.00410	0.00000496
1838	6	6	0.00737	0.00807	5	21	11	0.00367	0.00477
1	7	7	0.00737	0.00807	9 B	22	69	0.00360	0.00157
1829	8	8	0.00710	0.00764	3	23	18	0.00351	0.00349
47 B	9	9	0.00710	0.00764	250241 A	24	446	0.00346	0.00019
1830	10	10	0.00710	0.00764	499 B	25	129	0.00324	0.00105
648	11	4	0.00671	0.00894	163 B	26	184	0.00250	0.00068
11	12	5	0.00668	0.00890	22	27	24	0.00245	0.00279
2935 B	13	13	0.00451	0.00427	652	28	19	0.00241	0.00317
84 B	14	16	0.00451	0.00361	660	29	20	0.00241	0.00317
2934 B	15	14	0.00437	0.00425	80 B	30	49	0.00239	0.00194

ส่งผลให้ค่าคะแนนเพจแร็งค์ 30 ลำดับแรกหลังจากปรับลดผลกระทบของบูซฟาร์มในการคำนวณเพจแร็งค์จะได้ค่าคะแนนเพจแร็งค์ดังตารางที่ 4 ต่อไปนี้

ตารางที่ 4 คะแนนเพจแร็งค์ 30 ลำดับแรกหลังจากปรับลดลดผลกระทบของบูซฟาร์มในการคำนวณเพจแร็งค์

Rank	Page ID	Page Rank Score	Rank	Page ID	Page Rank Score
1	4	0.02624	16 B	84	0.00361
2	659	0.01651	17 A	250260	0.00357
3 A	250234	0.01378	18	3	0.00349
4	648	0.00894	19	652	0.00317
5	11	0.00890	20	660	0.00317
6	1838	0.00807	21	242	0.00316
7	1	0.00807	22	251	0.00316
8	1829	0.00764	23	7542	0.00284
9 B	47	0.00764	24	22	0.00279
10	1830	0.00764	25	240	0.00243
11	5	0.00477	26	272	0.00243
12 A	250240	0.00464	27	23	0.00241
13 B	2935	0.00427	28	7239	0.00234
14 B	2934	0.00425	29	292	0.00232
15 B	2	0.00425	30	9487	0.00220

จะเห็นได้จากผลการทดลองว่าหลังจากที่ใช้วิธีการที่นำเสนอปรับลดอิทธิพลของบูซฟาร์มเมื่อพิจารณาถึงเว็บเพจที่สแปมเพิ่มมา 3 กลุ่ม เว็บเพจเหล่านั้นจะถูกปรับลำดับไปยังลำดับที่ต่ำกว่า 30 อันดับแรกทั้งหมด ยกเว้นเว็บเพจที่ 250234, 250240 และ 250260 จากการพิจารณาถึงผลการทดลอง และโครงสร้างลิงค์โดยละเอียดเราพบว่า สาเหตุที่ทำให้เว็บเพจที่ 250234, 250240 และ 250260 ไม่ถูกปรับลดไปมากเหมือนกับเว็บเพจ อื่นๆที่ทำการสแปมเพิ่มนั้นเนื่องจากว่าเว็บเพจทั้ง 3

มีลิงค์จากเว็บเพจ 3, 5, 6, 15, 17, 22, 343, 652, 660, 3041 และ 3042 ซึ่งเป็นเว็บเพจที่มีคุณภาพ ส่งผลให้ของการปรับลดอิทธิพลที่นำเสนอลดลง อีกทั้งเว็บเพจทั้ง 3 สามารถถูกจัดในลำดับที่สูงกว่าเว็บเพจอื่นๆในบูซฟาร์ม

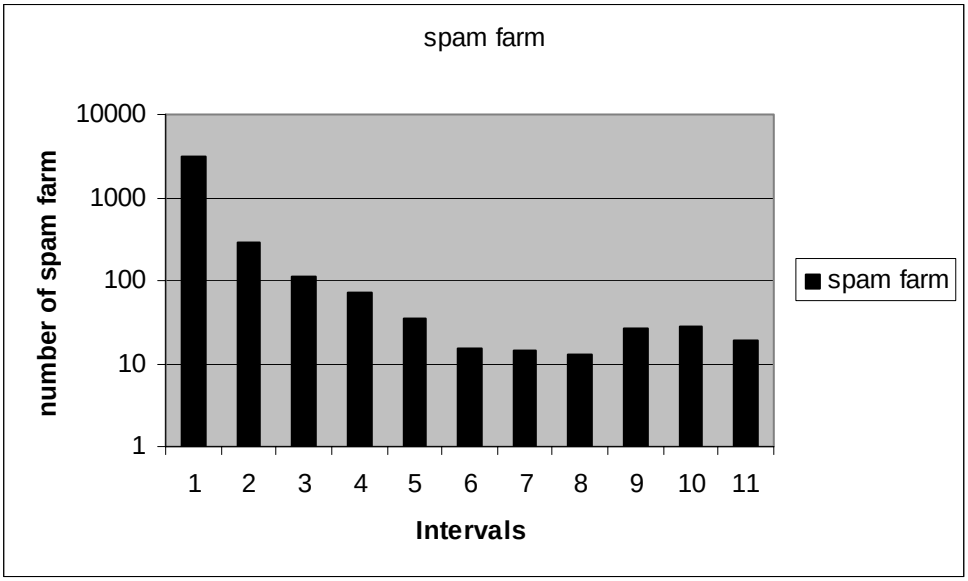
เมื่อพิจารณาถึงเว็บเพจที่ถูกค้นพบโดยวิธีการ (Wu and Brian, 2005a) ว่าเป็นสแปมฟาร์ม เว็บเพจเหล่านั้นจะถูกปรับลำดับไปยังลำดับที่ต่ำกว่า 30 อันดับแรกทั้งหมด ยกเว้นเว็บเพจที่ 2, 84, 2935 และ 2934 จากการพิจารณาถึงผลการทดลอง และโครงสร้างลิงค์โดยละเอียดเราพบว่าสาเหตุที่ทำให้เว็บเพจที่ 2, 84, 2935 และ 2934 ไม่ถูกปรับลดค่าคะแนนเพจแร็งค์ต่ำกว่า 30 ลำดับแรกเหมือนกับเว็บเพจอื่นๆ ที่ทำการสแปมเนื่องจากว่าเว็บเพจทั้งสามมีลิงค์จากเว็บเพจปรกติประมาณ 30 เว็บ และมีลิงค์ชี้มาจากเว็บเพจถูกตรวจพบว่าเป็นการสแปมโครงสร้างลิงค์ประมาณ 1-2 เว็บเพจ จึงสามารถล่าวได้ว่าค่าคะแนนเพจแร็งค์ของเว็บเพจทั้ง 3 นั้นไม่ขึ้นกับลิงค์ที่ชี้มาจากเว็บเพจที่เป็นสแปมโครงสร้างลิงค์ ดังนั้นเมื่อใช้วิธีการที่นำเสนอปรับลดอิทธิพลของบูซฟาร์มแล้วนั้น เว็บเพจทั้ง 3 สามารถถูกจัดลำดับใน 30 ลำดับแรกได้

จากตารางที่ 3 อาจมีข้อสงสัยว่าเหตุใดเว็บเพจในบูซฟาร์มบางเว็บเพจ อาทิเช่น เว็บเพจหมายเลข 250234 นั้นมีค่าคะแนนเพจแร็งค์สูงขึ้นเมื่อถูกนำมาปรับลดอิทธิพลของบูซฟาร์มที่นำเสนอ สาเหตุเนื่องจากค่าคะแนนของเว็บเพจนั้นขึ้นกับลิงค์จากเว็บเพจที่มีคุณภาพ 5 และ 15 เมื่อเว็บเพจทั้ง 2 ผ่านวิธีการที่นำเสนอแล้วนั้นมีค่าส่งผลให้มีค่าคะแนนสูงขึ้น จึงยังผลให้เว็บเพจหมายเลข 250234 มีค่าคะแนนเพจแร็งค์เพิ่มขึ้น

จากกรณีทั้ง 3 ที่ยกมาพิจารณาเราอาจจะกล่าวได้ว่าถ้าเว็บเพจใดๆในบูซฟาร์มที่ถูกชี้โดยเว็บเพจที่มีคุณภาพแล้วถูกนำมาปรับลดอิทธิพลของบูซฟาร์มด้วยวิธีการที่นำเสนอแล้ว เว็บเพจเหล่านั้นก็ยังสามารถถูกจัดลำดับในลำดับความสำคัญที่สูงได้ แต่ถ้าเป็นเว็บเพจใดๆในบูซฟาร์มที่ถูกลิงค์จากเว็บเพจที่มีคุณภาพต่ำจะถูกปรับลำดับความสำคัญให้อยู่ในลำดับที่เหมาะสม นั่นคือลำดับขึ้นกับผลรวมของค่าคะแนนเพจแร็งค์จากเว็บเพจคุณภาพต่ำเหล่านั้นส่งมาให้ เนื่องจากไม่สามารถสะสมค่าคะแนนเหล่านั้นเพื่อเพิ่มค่าคะแนนเพจแร็งค์ให้แก่ตนเองได้เนื่องจากวิธีการปรับลดอิทธิพลที่เราได้นำเสนอนั่นเอง

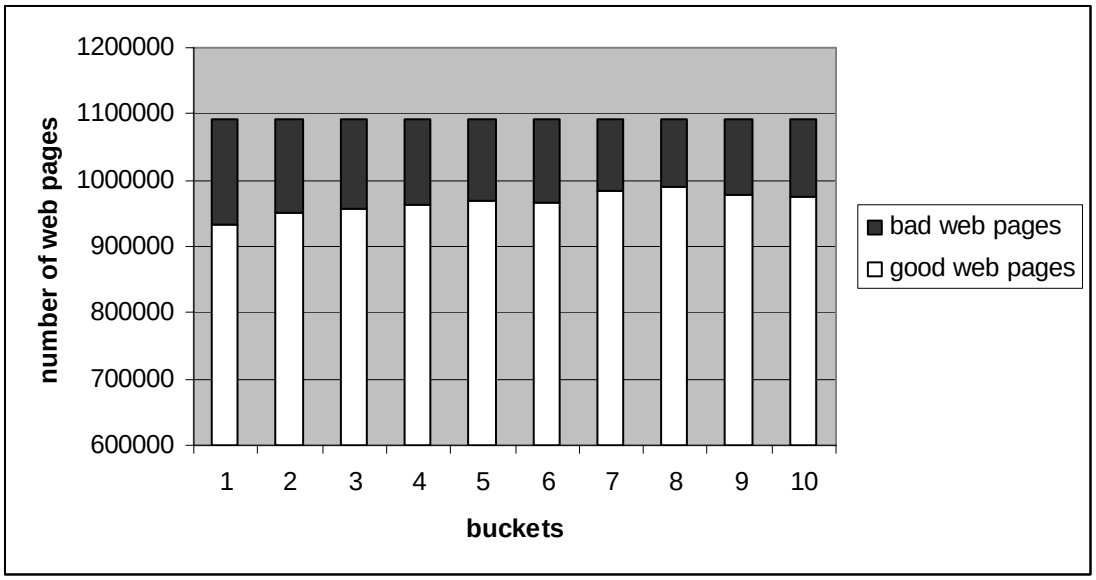
การทดลองบนเว็บกราฟในโดเมน ".th"

ในการวัดผลการทดลองในขั้นตอนนี้เรานำข้อมูลเว็บกราฟจากเว็บเพจในโดเมน ".th" โดยมีขนาด 10926864 เว็บเพจ ในขั้นตอนแรกนั้นเราจะทำการหาสแปมฟาร์มโดยใช้วิธีการ (Wu and Brian, 2005a) จากภาพที่ 34 แสดงถึงการกระจายตัวของขนาดเว็บเพจในสแปมฟาร์มทั้ง 11 ช่วง (interval) โดยช่วงแรกนั้นประกอบด้วยสแปมฟาร์มขนาด 1 – 10 ในช่วงที่สองประกอบด้วยสแปมฟาร์มขนาด 11 – 20 ตามลำดับ และช่วงที่ 11 นั้นประกอบด้วยสแปมฟาร์มขนาด 101 - ∞ จากการทดลองสแปมฟาร์มกลุ่มที่ใหญ่ที่สุดมีสมาชิก 1232323 เว็บเพจ และสแปมฟาร์มที่มีขนาดเล็กที่สุดมีขนาด 1 เว็บเพจ



ภาพที่ 34 กราฟแสดงการกระจายตัวของจำนวนเว็บเพจที่เป็นสมาชิกของสแปมฟาร์ม

หลังจากนั้นเราจะคำนวณค่าคะแนนเพจแร็งค์แก่ทุกเว็บเพจในเว็บกราฟ และจึงแบ่งช่วงค่าคะแนนเพจแร็งค์ออกเป็น 10 ช่วงด้วยกัน โดยแต่ละช่วงประกอบด้วยเว็บเพจจำนวนเท่าๆกันเรียงลำดับค่าคะแนนเพจแร็งค์จากมากไปน้อยตามลำดับ ยกตัวอย่าง เช่นในช่วงค่าคะแนนเพจแร็งค์ที่ 1 คือช่วงที่มีค่าเพจแร็งค์สูงสุด และในช่วงค่าคะแนนเพจแร็งค์ที่ 10 คือช่วงที่มีเพจแร็งค์ต่ำสุด จากภาพที่ 35 แสดงถึงการกระจายตัวของเว็บเพจในช่วงค่าคะแนนเพจแร็งค์ทั้ง 10 ช่วงโดยกราฟแท่งที่ลงสีทึบคือจำนวนเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์ซึ่งค้นพบจากอัลกอริทึม (Wu and Brian, 2005a) และกราฟแท่งที่ลงสีขาวคือจำนวนเว็บเพจที่ไม่เป็นการสแปมโครงสร้างลิงค์



ภาพที่ 35 การกระจายตัวของเว็บเพจใน 10 ช่วงค่าคะแนนเพจแร็งค์

ในช่วงการกระจายตัวของเว็บเพจทั้ง 10 เมื่อพิจารณาเฉพาะเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์นั้นประกอบด้วยคุณสมบัติที่น่าสนใจ 2 ชนิด นั่นคือ

1. จำนวนลิงค์เฉลี่ยต่อเว็บเพจที่ชี้มาจากเว็บเพจที่ดี(Average number of Good inLinks, *AGL*) กำหนดให้เซตของลิงค์จากเว็บเพจที่ดีที่ชี้มายังเว็บเพจ p แสดงด้วย L_p และจำนวนเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์ในช่วงการกระจายตัวของเว็บเพจ i แสดงด้วย N_i ดังนั้นเราจะสามารถแสดงสมการการคำนวณจำนวนลิงค์เฉลี่ยต่อเว็บเพจที่ชี้มาจากเว็บเพจที่ดีในช่วงกระจายตัวของเว็บเพจ i ได้ดังสมการต่อไปนี้

$$AGL_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\left|L_j\right| \tag{31}$$

2. ค่าคะแนนเพจแร็งค์เฉลี่ยของเว็บเพจที่ดีที่มีลิงค์ชี้มายังเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์ (Average PageRank of good supporter, *APR*) กำหนดให้เซตของเว็บเพจที่ไม่เป็นการสแปมโครงสร้างลิงค์ที่มีลิงค์ชี้มายังเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์ p แสดงด้วย G_p ดังนั้นเราจะ

สามารถเขียนสมการการคำนวณค่าคะแนนเพจแร็งค์เฉลี่ยต่อเว็บเพจเมื่อพิจารณาเว็บเพจที่ดีที่มีลิงค์ชี้มายังเว็บเพจที่เป็นการ สแปมโครงสร้างลิงค์ในช่วงกระจายตัวของเว็บเพจ i ได้ดังสมการต่อไปนี้

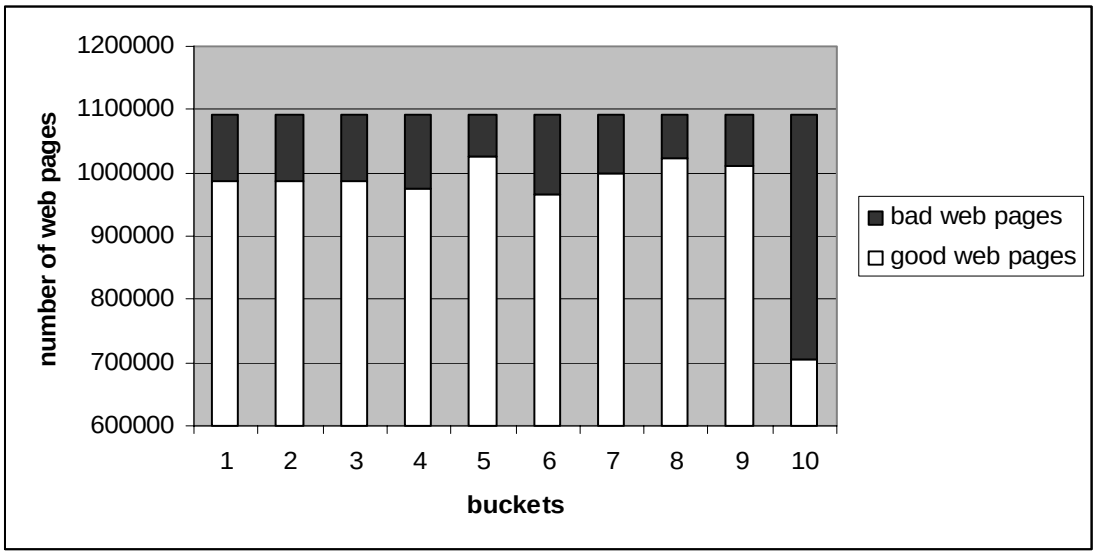
$$APR_i = \frac{\sum_{j=1}^{N_i} PageRank\ (G_j)}{\sum_{k=1}^{N_i} |L_k|} \qquad (32)$$

จากสมการที่ (31) และ (32) เราสามารถแสดงถึงคุณสมบัติทั้ง 2 ในช่วงการกระจายตัวของเว็บเพจทั้ง 10 ดังตารางที่ 5 ต่อไปนี้

ตารางที่ 5 จำนวนลิงค์เฉลี่ยที่ชี้มาจากเว็บเพจที่ดี และค่าคะแนนเพจแร็งค์เฉลี่ยของเว็บเพจที่ดีใน 10 ช่วงค่าคะแนนเพจแร็งค์

Bucket	AGL	APR
1	7.741381	1.06184E-7
2	0.843277	6.36128E-8
3	0.680782	7.50916E-8
4	0.517195	6.23083E-8
5	0.395006	6.48342E-8
6	0.361635	8.35578E-8
7	0.102205	1.54159E-7
8	0.108298	1.21968E-7
9	0.417236	1.13560E-7
10	0.825197	8.30951E-8

หลังจากนั้นได้นำวิธีการที่นำเสนอมาปรับลดอิทธิพลของบูซฟาร์ม แล้วไปคำนวณค่าเพจแร็งค์ตามสมการที่ (30) จะได้การกระจายตัวของเว็บเพจในช่วงค่าคะแนนเพจแร็งค์ต่างๆดังภาพที่ 36



ภาพที่ 36 การกระจายตัวของเว็บเพจใน 10 ช่วงค่าคะแนนเพจแร็งค์หลังจากปรับลดอิทธิพลของบูซฟาร์ม

และเมื่อใช้วิธีการที่นำเสนอมาปรับลดอิทธิพลของบูซฟาร์ม การกระจายตัวของเว็บเพจในช่วงทั้ง 10 เมื่อพิจารณาเฉพาะเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์นั้นสามารถแสดงคุณสมบัติทั้งสอง นั่นคือจำนวนลิงค์เฉลี่ยต่อเว็บเพจที่ชี้มาจากเว็บเพจที่ดี และค่าคะแนนเพจแร็งค์เฉลี่ยของเว็บเพจที่ดีที่มีลิงค์ชี้มายังเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์ ดังแสดงในตารางที่ 6 ต่อไปนี้

ตารางที่ 6 จำนวนลิงค์เฉลี่ยที่ชี้มาจากเว็บเพจที่ดี และค่าคะแนนเพจแร็งค์เฉลี่ยของเว็บเพจที่ดีใน 10 ช่วงค่าคะแนนเพจแร็งค์หลังจากการปรับลดผลกระทบของบูซฟาร์ม

Bucket	AGL	APR
1	11.287925	1.082621E-7
2	1.0811657	7.569986E-8
3	0.8902795	6.427950E-8
4	0.6053391	6.967266E-8
5	0.5842039	6.463632E-8
6	0.4209968	6.628511E-8
7	0.7080591	8.058620E-8
8	0.0366150	1.337882E-7
9	0.0087651	3.878108E-8
10	0.3875718	9.322711E-8

จะเห็นได้จากการทดลองว่าหลังจากนำวิธีการที่นำเสนอมาปรับลดอิทธิพลของบูซฟาร์มกับเว็บเพจที่ตรวจพบโดยวิธีการของ (Wu and Davidson, 2005a) ว่าเป็นเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์ พบว่าเว็บเพจบางส่วนที่ประกอบด้วยลิงค์คุณภาพต่ำนั้นถูกปรับลดความสำคัญ และถูกจัดอยู่ในช่วงค่าคะแนนเพจแร็งค์ที่ 10 จำนวนหนึ่ง และเมื่อพิจารณาถึงเว็บที่กระจายอยู่ตามช่วงค่าคะแนนเพจแร็งค์ จากการเปรียบเทียบระหว่างวิธีการที่นำเสนอและการคำนวณเพจแร็งค์ พบว่าการคำนวณเพจแร็งค์เมื่อพิจารณาคุณสมบัติทั้ง 2 ดังแสดงในตารางที่ 5 และ 6 การคำนวณเพจแร็งค์ไม่สามารถเรียงลำดับเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์ตามมาตรวัด AGL และ APR ซึ่งพบว่าการคำนวณเพจแร็งค์นั้นเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์สามารถถูกจัดลำดับในลำดับที่สูงโดยพิจารณาลิงค์จากเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์อื่นๆร่วมด้วย เนื่องจากลิงค์ที่ชี้มาจากเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์เราจะพิจารณาว่าเป็นลิงค์ที่ไม่น่าเชื่อถือ แตกต่างจากวิธีการที่นำเสนอนั้นสามารถจัดลำดับเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์ตามมาตรวัด AGL และ APR จึงสามารถกล่าวได้ว่าวิธีการที่นำเสนอนั้นพิจารณาลำดับของเว็บเพจที่เป็นสแปมโครงสร้างลิงค์โดยพิจารณาลิงค์ที่ชี้มาจากเว็บเพจที่ดีเป็นหลัก และไม่ขึ้นกับจำนวนลิงค์จากเว็บเพจที่เป็นสแปมโครงสร้างลิงค์อื่นๆ

วิจารณ์

จากผลการทดลอง แสดงให้เห็นว่าการสแปมโครงสร้างลิงค์มีผลกระทบต่ออัลกอริทึมเพจแร็งค์ (Page *et al*., 1998) โดยเว็บเพจที่ถูกสแปมโครงสร้างลิงค์นั้นสามารถถูกจัดลำดับใน 30 ลำดับแรกถึง 17 เว็บเพจ ซึ่งจัดอยู่ในลำดับที่ 2, 3, 4 9 13, 14, 15, 16, 17, 18, 19, 20, 22, 24, 25, 26 และ 30 ตามลำดับ ทำให้ผลค้นคืนที่ใช้การจัดลำดับด้วยอัลกอริทึมเพจแร็งค์ไม่น่าเชื่อถือ แต่เมื่อใช้วิธีการปรับลดอิทธิพลบูซฟาร์มในการคำนวณเพจแร็งค์แล้วนั้นปรากฏว่าสามารถลดจำนวนการสแปมโครงสร้างลิงค์เหลือเพียง 7 เว็บเพจคือเว็บเพจหมายเลขที่ 2, 84, 2934, 2935, 250234, 250240 และ 250260 สาเหตุที่ทำให้ทั้งสองเว็บเพจยังสามารถถูกจัดลำดับในลำดับที่สูง เนื่องจากทั้ง 7 เว็บเพจนี้มีลิงค์ชี้มาจากเว็บเพจที่คุณภาพจำนวนหนึ่ง ดังนั้นถึงแม้ว่าใช้วิธีการปรับลดอิทธิพลบูซฟาร์มในการคำนวณเพจแร็งค์ยังสามารถถูกจัดลำดับใน 30 ลำดับแรกได้

วิธีการที่นำเสนอนั้นจะทำการปรับลดผลกระทบเฉพาะลิงค์ที่ชี้ออกจากเว็บเพจที่ถูกค้นพบว่าเป็นการสแปมโครงสร้างลิงค์ และสร้างลิงค์เสมือนไปยังทุกเว็บเพจที่ไม่ใช่เว็บเพจในกลุ่มสแปมฟาร์มของตนเอง ซึ่งยังผลให้เว็บเพจที่อยู่ในกลุ่มสแปมฟาร์มนั้นมีแนวโน้มค่าคะแนนเพจแร็งค์ลดลง และเว็บเพจอื่นๆมีแนวโน้มค่าคะแนนเพจแร็งค์สูงขึ้น แต่อย่างไรก็ตามถ้าหากเว็บเพจใดๆที่ไม่เป็นสมาชิกสแปมฟาร์มนั้นๆ แต่ค่าคะแนนเพจแร็งค์ขึ้นกับลิงค์ที่ชี้มาจากเว็บเพจในสแปมฟาร์มก็จะส่งผลให้เว็บเพจนั้นมีค่าเพจแร็งค์ลดลง จากกรณีเหล่านี้เราอาจทำการกำหนดให้เว็บเพจที่กล่าวมาเป็นส่วนหนึ่งของสแปมฟาร์มเนื่องจากค่าเพจแร็งค์ขึ้นกับลิงค์ที่ชี้มาจากเว็บเพจในสแปมฟาร์มนั่นเอง

ดังนั้นเราสามารถกล่าวได้ว่าวิธีการปรับลดอิทธิพลบูซฟาร์มในการคำนวณเพจแร็งค์นี้สามารถปรับลดผลกระทบของการสแปมโครงสร้างลิงค์ที่ไม่มีลิงค์ชี้มาจากเว็บเพจที่มีคุณภาพหรือมีลิงค์ชี้จากเว็บเพจที่มีคุณภาพจำนวนน้อย และวิธีการที่นำเสนอนี้มีผลกระทบน้อยต่อโครงสร้างลิงค์ที่มีโครงสร้างคล้ายคลึงกับการสแปมโครงสร้างลิงค์ที่มีลิงค์ชี้จากเว็บเพจที่มีคุณภาพจำนวนมาก แต่อย่างไรก็ตามวิธีการที่นำเสนอนั้นไม่สามารถปรับลดอิทธิพลของการสแปมโครงสร้างลิงค์ที่สามารถทำให้เว็บเพจของตนจัดอยู่ในลำดับที่สูงจากลิงค์ที่ชี้มาจากเว็บเพจที่ไม่ถูกตรวจพบว่าเป็นการสแปมโครงสร้างลิงค์ ซึ่งจากผลการทดลองเว็บเพจหมายเลขที่ 2, 84, 2934 และ 2935 นั้นหลังจากนำวิธีการที่นำเสนอมาปรับลดอิทธิพลบูซฟาร์มแล้วยังถูกจัดลำดับใน 30 ลำดับแรก เมื่อพิจารณาลิงค์ชี้เข้าประมาณ 30 ลิงค์ที่ชี้เข้าเว็บเพจทั้ง 4 นั้นเป็นลิงค์ที่ชี้มาจากเว็บเพจที่ทำ

การขโมยลิงค์ หรือจากการซื้อลิงค์จากผู้พัฒนา 30 เว็บไซต์เหล่านั้นมา การสแปมโครงสร้างลิงค์ลักษณะนี้ก็ยังจะส่งผลกระทบต่อวิธีการที่นำเสนอ แต่ในความเป็นจริงนั้นเป็นการยากที่จะสามารถรวบรวมลิงค์คุณภาพจำนวนมากจากการการขโมยลิงค์ หรือจากการซื้อลิงค์จากเว็บไซต์ที่มีคุณภาพและเว็บเพจที่มีคุณภาพนั้นน้อยครั้งที่จะชี้ไปยังเว็บเพจที่ทำการสแปมโครงสร้างลิงค์จึงสามารถเชื่อได้ว่าวิธีการนี้เหมาะสมกับการนำไปใช้กับการปรับลดผลกระทบการสแปมโครงสร้างลิงค์ของเว็บกราฟที่เก็บรวบรวมจากอินเทอร์เน็ตได้

สรุปและข้อเสนอแนะ

สรุป

วิธีการปรับลดอิทธิพลของบูซฟาร์มในการคำนวณเพจแร็งค์เป็นวิธีการที่พัฒนาเพื่อให้สามารถนำไปใช้กับการคำนวณเพจแร็งค์ได้ โดยการสร้างเมตริกซ์ชนิดใหม่อีก 3 ชนิดด้วยกัน เพื่อแบ่งเว็บกราฟออกเป็น 3 ส่วน หลังจากนั้นจึงนำเมตริกซ์ทั้ง 3 มาสร้างเมตริกซ์ความสัมพันธ์ใหม่ที่ทำการปรับผลกระทบของบูซฟาร์ม โดยเมตริกซ์ความสัมพันธ์ใหม่ที่ทำการปรับผลกระทบของบูซฟาร์มนั้นยังคงคล้อยตามคุณสมบัติของมาคอร์ฟ และเมื่อเมตริกซ์ความสัมพันธ์มีคุณสมบัติดังกล่าวแล้ว การคำนวณแบบวนซ้ำในสมการที่ (30) จะรับประกันได้ว่าเมื่อจำนวนรอบในการคำนวณ $i = \infty$ แล้วได้คำตอบของแร็งค์เวกเตอร์ลู่เข้าเสมอ

ในงานวิจัยฉบับนี้มีการคำนวณหาอัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์มเพื่อนำไปใช้ปรับน้ำหนักลิงค์เสมือนที่กระจายค่าเพจแร็งค์ผ่านการสร้างลิงค์เสมือนให้แก่เว็บเพจอื่นนอกจากบูซฟาร์มกลุ่มนั้นๆ ซึ่งผลจากการทดลองเบื้องต้นแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถปรับลดค่าคะแนนเพจแร็งค์เว็บเพจที่อยู่ในกลุ่มของบูซฟาร์มได้จริง โดยการปรับลดนั้นจะขึ้นอยู่กับโครงสร้างลิงค์ของกลุ่มบูซฟาร์มนั้นๆ โดยที่ไม่พิจารณาปรับลดลิงค์ที่ชี้ยังโครงสร้างลิงค์ของบูซฟาร์มที่เราพิจารณาอยู่ ทำให้หลังจากปรับลดอิทธิพลของบูซฟาร์มแล้วลำดับเว็บเพจภายในบูซฟาร์มยังสอดคล้องสมมติฐานของเพจแร็งค์ดั้งเดิม นั่นคือ ลิงค์ที่ชี้มายังเว็บเพจที่พิจารณาสามารถแสดงถึงความสำคัญของเว็บเพจเหล่านั้น เป็นผลทำให้วิธีการนี้เหมาะสมต่อการนำไปใช้งานกับการปรับลดและจัดลำดับเว็บเพจบนอินเทอร์เน็ตด้วยอัลกอลิทึมเพจแร็งค์ได้เป็นอย่างดี

ในวิทยานิพนธ์เล่มนี้ได้ชี้ให้เห็นถึงปัญหาของการสแปมโครงสร้างลิงค์ในการจัดลำดับด้วยอัลกอลิทึมเพจแร็งค์ โดยวิธีการที่นำเสนอใหม่ คือการคำนวณหาอัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์ม และการสร้างโครงสร้างลิงค์เสมือนที่ทำหน้าที่ปรับลดผลกระทบของบูซฟาร์มโดยกระจายค่าคะแนนเพจแร็งค์ของบูซฟาร์มนั้นให้กับทุกๆเว็บเพจที่ไม่ใช่เว็บเพจในบูซฟาร์มกลุ่มนั้นๆ

ข้อเสนอแนะ

เพื่อผลลัพธ์ที่ถูกต้องของวิธีการปรับลดอิทธิพลของบูซฟาร์มในการคำนวณเพจแร็งค์นั้น ในขั้นตอนการค้นหาบูซฟาร์มนั้นจะต้องหาวิธีการที่สามารถค้นหาบูซฟาร์มมาได้ครบถ้วน เนื่องจากวิธีการที่นำเสนอนี้ไม่สามารถให้ผลลัพธ์ให้การจัดลำดับที่ถูกต้องได้เมื่อ โครงสร้างของบูซฟาร์มที่ค้นคืนกลับมานั้นไม่ครบถ้วน

ในส่วนของการจัดกลุ่มของบูซฟาร์มก่อนการคำนวณค่าอัตราการเปลี่ยนแปลงของความน่าจะเป็นเฉลี่ยของบูซฟาร์มนั้นต้องมีการเพิ่มส่วนในการพิจารณาให้มีความละเอียดมากยิ่งขึ้น โดยจะต้องคำนึงถึงการคำนวณอัตราการเปลี่ยนแปลงของความน่าจะเป็นเฉลี่ยของบูซฟาร์ม ซึ่งในงานวิจัยที่นำเสนอนี้พิจารณาเพียงแต่เมื่อเว็บเพจใดๆที่ถูกกำหนดเป็นการสแปมโครงสร้างลิงค์แล้วมีลิงค์ชี้มายังหรือถูกชี้จากกลุ่มบูซฟาร์มที่พิจารณาก็จะถูกกำหนดว่าเว็บเพจนั้นเป็นส่วนหนึ่งของกลุ่มบูซฟาร์มที่พิจารณาอยู่ ซึ่งอาจทำให้โครงสร้างของบูซฟาร์มที่ค้นคืนกลับมานั้นไม่ครบถ้วนหรือมีบางส่วนของบูซฟาร์มกลุ่มอื่นมาปรากฏในกลุ่มบูซฟาร์มที่เราพิจารณาอยู่

และเพื่อความถูกต้องในขั้นตอนการคำนวณค่าอัตราการเปลี่ยนแปลงของค่าความน่าจะเป็นเฉลี่ยของบูซฟาร์มแต่ละกลุ่มนั้นเมื่อพิจารณาเว็บกราฟหลักมีขนาดแตกต่างกัน เราอาจพิจารณาการสุ่มกระโดดไปจากเว็บเพจใดๆไปยังเว็บเพจ x แทนด้วยการสุ่มกระโดดจากเว็บเพจใดๆไปยังเว็บเพจทั้งหมดในเว็บกราฟยกเว้นเว็บเพจที่เป็นสมาชิกของกลุ่มบูซฟาร์มที่เราพิจารณาอยู่ และพิจารณาค่า N ที่ใช้คำนวณในเว็บกราฟจำลองใดๆมีค่าเท่ากับจำนวนเว็บเพจทั้งหมดในเว็บกราฟหลัก

เอกสารและสิ่งอ้างอิง

Alan, P. 2001. **The Classification of Search Engine Spam.** SE Spam Classification. Available Source: http://www.silverdisc.co.uk/articles/spam-classification/, May 1, 2007.

Baeza-Yates, R., P. Boldi and C. Castillo. 2006. Generalizing pagerank: Damping functions for link-based ranking algorithms, pp. 308-315. *In* **Proceedings of ACM SIGIR** . Seattle, Washington, USA.

Becchetti, L., C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. 2006. Using rank propagation and probabilistic counting for link-based spam detection, *In* **Technical report** . Dynamically Evolving, Large-Scale Information Systems(DELIS).

Benczur, A.A., K. Csalogany, T. Sarlos and M. Uher. 2005. Spamrank: fully automatic link spam detection, *In* **Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web** . Chiba, Japan.

Bharat, K. and M. Henzinger. 1998. Improved algorithms for topic distillation in hyperlinked environments, pp. 104-111. *In* **Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)** . Melbourne, Australia.

Cohen, E. 1997. Size-estimation framework with application to transitive closure and reachability, pp. 441-453. *In* **Journal of Computer and System Sciences** .

Flajolet, P. and N.G. Martin. 1985. Proabilistic counting algorithms for data base applications, pp. 182-209. *In* **Journal of Computer and System Sciences** .

Fogaras, D. and B. Racz. 2004. Towards scaling fully personalized PageRank, น. 105-117. *ใน* **Proceeding of the 3rd Workshop on Algorithm and Models for the Web-Graph (WAW)** . Rome, Italy.

Google. Available Source: http://www.google.com/, May 1, 2007.

Gyongyi, Z. and H. Garcia-Molina. 2004. Combating web spam with TrustRank, pp. 576-587. *In* **Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)**. Toronto, Canada.

_____ and _____ 2005a. Web spam taxonomy, *In* **Proceedings of 1st International Workshop on Adversarial Information Retrieval on the Web(AirWeb)**. Chiba, Japan.

_____ and _____ 2005b. Link Spam Alliances, pp. 517-528. *In* **Proceedings of the 31th International Conference on Very Large Data Bases (VLDB)**. Trondheim, Norway.

Henzinger, M., R. Motwani and C. Silverstein. 2002. Challenges in web search engines, pp. 1573-1579. *In* **Proceedings of International Joint Conference on Artificial Intelligence.** Chiba, Japan.

Jeh, G. and J. Widom. 2003. Scaling personalized web search, pp. 271-279. *In* **Proceedings of the Twelfth International World Wide Web Conference** . Honolulu, Hawaii, USA.

Kleinberg, J.M. 1999. Authoritative source in a hyperlinked environment, pp. 604-632. *In* **Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA-98)** . San Francisco, CA.

Leonard Kleinrock. 1975. **QUEUEING SYSTEMS VOLUME I: THEROY.** 1 ed. A Wiley-Interscience Publication

Markus, S. 2003. **PR0 - Google's PageRank 0 Penalty.** Available Source: http://pr.efactory.de/e-pr0.shtml, May 1, 2007.

Page, L., S. Brin, R. Motwani and T. Winogard. 1998. The pagerank citation ranking: Bringing order to the web, *In* **Technical report**. Stanford Digital Library Technologies Project.

Wu, B. and D.B. Davidson. 2005a. Identifying Link Farm Spam Pages, pp. 820-829. *In* **Proceedings of the 14th International World Wide Web Conference 2005**. Chiba, Japan.

_____ and _____ 2005b. Undue influence: Eliminating the impact of link plagiarism on web search rankings, *In* **Technical report**. LeHigh University.

Yahoo. Available Source: http://www.yahoo.com/, August 1, 2006.

ภาคผนวก

ภาคผนวก ก

บทพิสูจน์ค่า ACB_i มีขอบเขต [0,1]

บทพิสูจน์ค่า ACB_i มีขอบเขต [0,1]

เมื่อพิจารณาถึงค่า $Pr(i_n)$ (ผลรวมค่าคะแนนเพจแร็งค์ของเว็บเพจที่เป็นสมาชิกของบูซฟาร์ม i ใดๆในเว็บกราฟจำลอง) จะประกอบด้วยค่าเพจแร็งค์จาก 3 ส่วนด้วยกันนั่นคือ

1. ค่าเพจแร็งค์จากการสุ่มกระโดด(randomjump) เมื่อพิจารณาว่าอยู่ ณ เว็บเพจ X ไปยังทุกๆเว็บเพจที่เป็นสมาชิกของบูซฟาร์ม i ใดๆ

2. ค่าสุ่มกระโดดเมื่อพิจารณาว่าอยู่ ณ เว็บเพจที่เป็นสมาชิกของบูซฟาร์ม i ใดๆไปยังทุกๆเว็บเพจที่เป็นสมาชิกของบูซฟาร์ม i ใดๆ

3. ค่าเพจแร็งค์จากเว็บเพจที่มีลิงค์ชี้ไปยังเว็บเพจที่เป็นสมาชิกของบูซฟาร์ม i ใดๆ

$$\begin{aligned}\Pr(i_n) &= \frac{N-1}{N}(1-c)PageRank(X_{n-1}) + \frac{N-1}{N}(1-c)\Pr(i_{n-1}) + \\ &\quad c\sum_{j\in B(BF_i)}\frac{1}{\omega(j)}PageRank(j_{n-1}) \\ \Pr(i_n) &= RandonJump(i_n) + \frac{N-1}{N}(1-c)\Pr(i_{n-1}) + \\ &\quad c\sum_{j\in B(BF_i)}\frac{1}{\omega(j)}PageRank(j_{n-1})\end{aligned} \tag{33}$$

เมื่อพิจารณาถึงค่าเพจแร็งค์จากการสุ่มกระโดด เมื่อพิจารณาว่าอยู่ ณ เว็บเพจ X ไปยังทุกๆเว็บเพจที่เป็นสมาชิกของบูซฟาร์ม i ใดๆ จะสามารถแสดงได้ว่า

1. $PageRank(X_{n-1})$ มีขอบเขต [0,1] เนื่องจากเป็นค่าความน่าจะเป็น

2. $\frac{N-1}{N}$ มีขอบเขต [0.5,1] เนื่องจาก N มีขอบเขต [2, ∞] และ

3. $(1-c)$ มีขอบเขต [0,1] เนื่องจากค่า α มีขอบเขต [0,1]

ดังนั้นค่า $\frac{N-1}{N}(1-c)PageRank(X_{n-1})$ มีขอบเขต [0,1]

เมื่อพิจารณาถึงค่าสุ่มกระโดดเมื่อพิจารณาว่าอยู่ ณ เว็บเพจที่เป็นสมาชิกของบูซฟาร์ม i ใดๆไปยังทุกๆเว็บเพจที่เป็นสมาชิกของบูซฟาร์ม i ใดๆ จะสามารถแสดงได้ว่า

1. $\Pr(i_{n-1})$ มีขอบเขต [0,1] เนื่องจากเป็นผลรวมของค่าความน่าจะเป็นที่ผู้ใช้งานที่เดินทางตามลิงค์ด้วยความน่าจะเป็นเท่ากันจะอยู่ ณ เว็บเพจนั้นๆจำนวน *N-1* เว็บเพจ เมื่อผลรวมของความน่าจะเป็นที่ผู้ใช้งานที่เดินทางตามลิงค์ด้วยความน่าจะเป็นเท่ากันจะอยู่ ณ เว็บเพจนั้นๆ จำนวน N เว็บเพจ มีค่าเท่ากับ 1 เมื่อกำหนดให้ N คือจำนวนเว็บเพจทั้งหมดในเว็บกราฟจำลอง

2. $\frac{N-1}{N}$ มีขอบเขต [0.5, 1] เนื่องจาก N มีขอบเขต [2, ∞] และ

3. $(1-c)$ มีขอบเขต [0, 1]

เมื่อพิจารณาถึงค่าเพจแร็งค์จากเว็บเพจที่มีลิงค์ชี้ไปยังเว็บเพจที่เป็นสมาชิกของบูซฟาร์ม i ใดๆ จะสามารถแสดงได้ว่า

1. $\sum_{j\in B(BF_i)} PageRank(j_{n-1})$ มีขอบเขต [0,1] เนื่องจาก $B(BF_i)\subset BF_i$ และผลรวมของเว็บเพจที่เป็นสมาชิกของบูซฟาร์ม i (BF_i)คือ $\Pr(i_n)$ ซึ่งมีขอบเขต [0,1]

2. $\frac{1}{\omega(j)}$ มีขอบเขต [0,1] เมื่อ $\omega(j)$ คือจำนวนลิงค์ที่ชี้ออกจากเว็บเพจ j ใด โดยที่ค่า $\frac{1}{\omega(j)}$ จะมีค่าเท่ากับ 0 เมื่อจำนวนลิงค์ที่ชี้ออกจากเว็บเพจ j ใดมีค่าเท่ากับ ∞ และมีค่าเท่ากับ 1 เมื่อจำนวนลิงค์ที่ชี้ออกจากเว็บเพจ j ใดมีค่าเท่ากับ 1

3. c มีขอบเขต [0, 1] เนื่องจากเป็นค่าที่กำหนดโดยผู้วิจัยเอง โดยในการคำนวณเพจแร็งค์นั้นกำหนดให้มีค่าเท่ากับ 0.85

ดังนั้นค่า $c \sum_{j \in B(BF_i)} \frac{1}{\omega(j)} PageRank(j_{n-1})$ มีขอบเขต [0, 1]

จากสมการที่ (33) และขอบเขตทั้ง 2 ของพจน์ที่กล่าวสามารถแสดงได้ว่า

$$\Pr(i_n) \geq RandomJump(i_n) \tag{34}$$

จากสมการที่ (24) ค่า ACB_i สามารถแสดงได้ดังสมการต่อไปนี้

$$ACB_i = \frac{1}{K_i} \sum_{n=1}^{K_i} (\Pr(i_{n-1}) - (\Pr(i_n) - Randomjump(i_n))) / \Pr(i_{n-1})$$

กำหนดให้ $\Pr(i_n) - Randomjump(i_n)$ แทนด้วยตัวแปร c โดยขอบเขตของ $\Pr(i_n)$ และ $RandomJump(i_n)$ มีขอบเขตระหว่าง [0,1] และจากสมการที่ (32) สามารถบอกได้ว่า c มีขอบเขต [0,1]

จากสมการที่ (33) สามารถแสดงได้ดังต่อไปนี้

$$\Pr(i_n) - RandonJump(i_n) = \frac{N-1}{N}(1-c)\Pr(i_{n-1}) + c \sum_{j \in B(BF_i)} \frac{1}{\omega(j)} PageRank(j_{n-1})$$

$$z = \frac{N-1}{N}(1-\alpha)\Pr(i_{n-1}) + \alpha \sum_{j \in B(BF_i)} \frac{1}{\omega(j)} PageRank(j_{n-1})$$

เพื่อจะพิสูจน์ว่า $\Pr(i_{n-1}) \geq k$ จะพิจารณาในกรณีค่า $c \sum_{j \in B(BF_i)} \frac{1}{\omega(j)} PageRank(j_{n-1})$ มีค่าสูงสุด เมื่อทุกๆเว็บเพจที่เป็นสมาชิกของบูซฟาร์ม i ใดๆมีลิงค์ชี้แต่เว็บเพจในกลุ่มของตน (BF_i) ซึ่งจะสามารถแสดงได้ดังต่อไปนี้

$$z = \frac{N-1}{N}(1-c)\Pr(i_{n-1}) + c\Pr(i_{n-1})$$

$$z = (\frac{N-1}{N}(1-c) + c)\Pr(i_{n-1})$$

พิจารณาพจน์ $1-c$ และ c มีขอบเขต [0, 1] และ $1-c$ เป็นส่วนกลับของ c ดังนั้น

$$(1-c)+c=1$$

เมื่อ $\frac{N-1}{N}$ ซึ่งมีขอบเขต [0.5,1] แล้ว $\frac{N-1}{N}(1-c)+c$ มีขอบเขต [0.5, 1] เสมอ

$$\Pr(i_{n-1}) \geq z \tag{35}$$

จากสมการที่ (34) จะสามารถแสดงสมการการคำนวณค่า ACB_i ดังต่อไปนี้

$$ACB_i = \frac{1}{K_i}\sum_{n=1}^{K_i}(\Pr(i_{n-1})-z)/\Pr(i_{n-1})$$
$$ACB_i = \frac{1}{K_i}\sum_{n=1}^{K_i}(1-\frac{z}{\Pr(i_{n-1})})$$

กำหนดให้ $\frac{z}{\Pr(i_{n-1})}$ แทนด้วยตัวแปร d และจากสมการที่ (35) จะสามารถแสดงได้ว่า d มีขอบเขต [0, 1] ดังนั้นจึงสรุปได้ว่าค่า ACB_i จะมีขอบเขต [0, 1] เสมอ

ภาคผนวก ข

กรณีศึกษา $\left|BF_1\right| + \left|BF_2\right| \approx N$

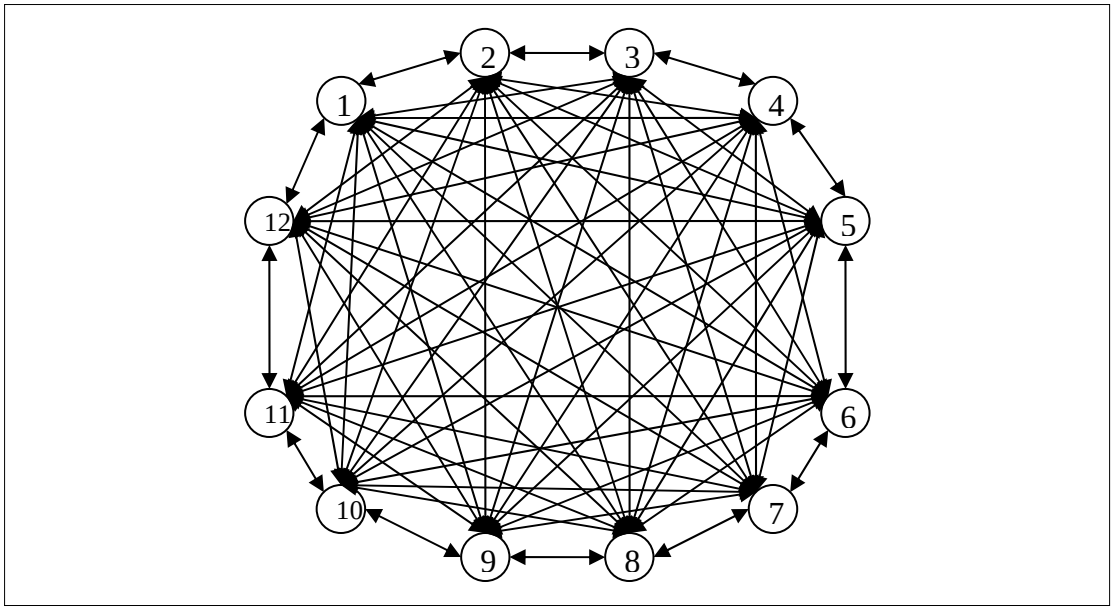
กรณีศึกษา $|BF_1| + |BF_2| \approx N$

เมื่อกำหนดให้ N คือจำนวนเว็บเพจในเว็บกราฟทั้งหมด และ BF_i คือเซตของเว็บเพจที่เป็นสมาชิกของบูซฟาร์มกลุ่มที่ i พิจารณาเว็บกราฟที่ประกอบด้วยบูซฟาร์มจำนวน 2 กลุ่ม และ จำนวนเว็บเพจของบูซฟาร์มทั้งสองกลุ่มรวมกันแล้วมีจำนวนใกล้เคียงเว็บเพจทั้งหมดในเว็บกราฟดังแสดงได้ดังสมการต่อไปนี้

$$|BF_1| + |BF_2| \approx N$$

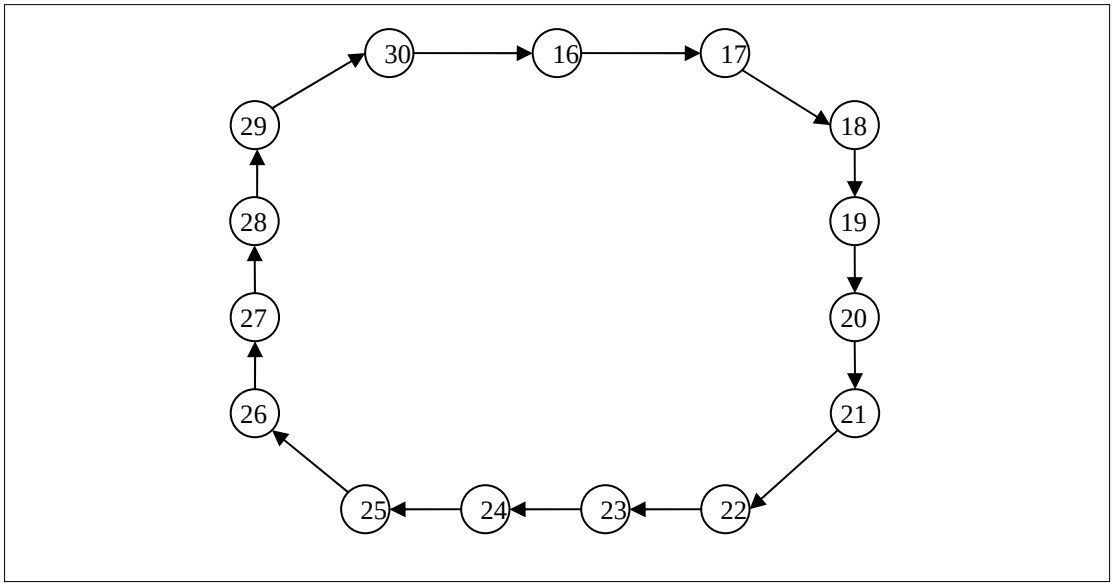
ซึ่งในกรณีศึกษานี้จะขอยกตัวอย่างเว็บกราฟขนาด 30 เว็บเพจ ซึ่งประกอบด้วยการสแปมโครงสร้างลิงค์ หรือบูซฟาร์มจำนวน 2 กลุ่มคือ

1. เว็บโครงสร้างสมบูรณ์ ซึ่งมีลักษณะคือทุกเว็บเพจในเว็บโครงสร้างสมบูรณ์จะมีลิงค์ชี้ไปยังทุกๆเว็บเพจในกลุ่มเว็บโครงสร้างสมบูรณ์ โดยกำหนดให้มีขนาด 12 เว็บเพจ ประกอบด้วยเว็บเพจหมายเลข 1 – 12 สามารถแสดงได้ดังภาพต่อไปนี้



ภาพผนวกที่ ข1 เว็บโครงสร้างสมบูรณ์ในกรณีศึกษา

2. เว็บวงแหวน ซึ่งมีลักษณะโครงสร้างลิงค์คล้ายคลึงกับวงแหวน โดยกำหนดให้มีขนาด 15 เว็บเพจ ประกอบด้วยเว็บเพจหมายเลข 16 - 30 สามารถแสดงได้ดังภาพต่อไปนี้



ภาพผนวกที่ ข2 เว็บวงแหวนในกรณีศึกษา

จากโครงสร้างการสแปมโครงสร้างลิงค์ทั้ง 2 แบบนั้นเราจะออกแบบโครงสร้างเว็บกราฟในกรณีศึกษาโดยมีเว็บเพจ 13, 14 และ 15 เป็นตัวแบ่งโครงสร้างการสแปมโครงสร้างลิงค์ออกจากกัน เมื่อพิจาณาเว็บเพจหมายเลข 13 จะมีลิงค์ชี้ไปยังเว็บเพจหมายเลข 3 และถูกลิงค์มาจากเว็บเพจหมายเลข 29 เมื่อพิจาณาเว็บเพจหมายเลข 14 จะมีลิงค์ชี้ไปยังเว็บเพจหมายเลข 3, 15, 18 และถูกลิงค์มาจากเว็บเพจหมายเลข 13 และ 22 เมื่อพิจารณาเว็บเพจหมายเลข 15 นั้นจะมีลิงชี้ไปยังเว็บเพจหมายเลข 18 และถูกลิงค์มาจากเว็บเพจหมายเลข 11 และ 14

ขั้นตอนต่อไปเราจะทำการคำนวณค่าคะแนนเพจแร็งค์ให้แก่ทุกๆเว็บเพจในเว็บกราฟกรณีศึกษา ซึ่งสามารถแสดงได้ดังตารางผนวกที่ ข1 โดยแถวที่มี (A) ในลำดับคือเว็บเพจที่กำหนดให้เป็นเว็บเพจที่ทำการสแปมโครงสร้างลิงค์หรือบูซฟาร์มนั่นเอง

ตารางผนวกที่ ข1 คะแนนเพจแร็งค์ของเว็บกราฟในกรณีศึกษา

Rank	Page ID	Page Rank Score	Rank	Page ID	Page Rank Score
1 A	3	0.05490	16 A	11	0.03927
2 A	18	0.04521	17 A	12	0.03927
3 A	19	0.04343	18	14	0.02917
4 A	20	0.04192	19 A	29	0.02899
5 A	21	0.04063	20 A	28	0.02822
6 A	22	0.03955	21 A	27	0.02731
7 A	1	0.03927	22 A	26	0.02625
8 A	2	0.03927	23 A	25	0.02500
9 A	4	0.03927	24 A	24	0.02353
10 A	5	0.03927	25 A	23	0.02180
11 A	6	0.03927	26 A	17	0.02176
12 A	7	0.03927	27 A	16	0.01972
13 A	8	0.03927	28 A	30	0.01732
14 A	9	0.03927	29	13	0.01732
15 A	10	0.03927	30	15	0.01583

หลังจากนั้นได้นำวิธีการที่นำเสนอมาปรับลดอิทธิพลของบูซฟาร์ม แล้วไปคำนวณค่าเพจแร็งค์ตามสมการที่ (30) จะได้ค่าคะแนนเพจแร็งค์ดังตารางผนวกที่ ข2 ต่อไปนี้

ตารางผนวกที่ ข2 คะแนนเพจแร็งค์ของเว็บกราฟในกรณีศึกษาหลังจากปรับลดลดผลกระทบของบูซฟาร์มในการคำนวณเพจแร็งค์

Rank	Page ID	Page Rank Score	Rank	Page ID	Page Rank Score
1 A	18	0.09800	16 A	11	0.02854
2 A	3	0.06713	17 A	12	0.02854
3	14	0.06636	18 A	20	0.02459
4	15	0.06461	19 A	21	0.02429
5	13	0.04657	20 A	22	0.02428
6 A	19	0.02909	21 A	29	0.024278
7 A	1	0.02854	22 A	28	0.024278
8 A	2	0.02854	23 A	27	0.024278
9 A	4	0.02854	24 A	26	0.024278
10 A	5	0.02854	25 A	25	0.024275
11 A	6	0.02854	26 A	17	0.024275
12 A	7	0.02854	27 A	24	0.02422
13 A	8	0.02854	28 A	16	0.02422
14 A	9	0.02854	29 A	23	0.02348
15 A	10	0.02854	30 A	30	0.02348

จะเห็นได้จากผลการทดลองพบว่าหลังจากที่ใช้วิธีการที่นำเสนอสามารถปรับลดอิทธิพลของบูซฟาร์มกับเว็บกราฟในกรณีศึกษาแล้วนั้นเว็บเพจหมายเลข 13, 14 และ 15 สามารถถูกจัดอยู่ในลำดับที่สูงคือลำดับที่ 5, 3 และ 4 ตามลำดับซึ่งขึ้นมาจากลำดับ 29, 18 และ 30 ในการคำนวณเพจแร็งค์และเว็บเพจหมายเลข 18 และ 3 ซึ่งเป็นเว็บเพจที่กำหนดให้เป็นสมาชิกเป็นเว็บเพจที่ทำการสแปมโครงสร้างลิงค์หรือบูซฟาร์มยังสามารถถูกจัดอยู่ในลำดับที่สูงคือลำดับที่ 1 และ 2 เพราะว่ามีลิงค์ชี้มาจากเว็บเพจหมายเลข 13, 14 และ 15 ซึ่งเราอาจกล่าวได้ว่าเป็นลิงค์ที่ชี้มาจากเว็บเพจที่มีคุณภาพ ซึ่งจากการทดลองขนาดเล็กๆนี้นั่นเองทำให้เชื่อได้ว่าวิธีการที่นำเสนอสามารถแก้ไขเว็บกราฟที่ประกอบด้วยบูซฟาร์มจำนวน 2 กลุ่ม และจำนวนเว็บเพจของบูซฟาร์มทั้งสองกลุ่มรวมกันแล้วมี

จำนวนใกล้เคียงเว็บเพจทั้งหมดในเว็บกราฟได้ แต่อย่างไรก็ตามในกรณีที่ทั้งเว็บกราฟเป็นเว็บเพจทั้งหมดเป็นการสแปมโครงสร้างลิงค์วิธีการที่นำเสนอจะเกิดข้อผิดพลาดยังผลให้ค่าคะแนนเพจแร็งค์ทั้งหมดมีค่าเท่ากับ 0 แต่ไม่มีผลกับกรณีที่ทั้งเว็บกราฟเป็นเว็บเพจปกติทั้งหมด

ภาคผนวก ค

คำอธิบายตัวแปรที่ใช้ในงานวิจัย

คำอธิบายตัวแปรที่ใช้ในงานวิจัย

$\omega(u)$	=	จำนวนลิงค์ที่ชี้ออกจากเว็บเพจ u
ρ	=	ค่า correlation coefficient ระหว่างการกระจายตัวแบบพาวเวอร์ลอร์และค่าคะแนนเพจแร็งค์ของเว็บเพจช่วยเหลือของเว็บเพจที่พิจารณา
ρ_0	=	ค่ามากที่สุดที่ยอมรับได้ที่ใช้พิจารณาค่า ρ
ε	=	เซตของลิงค์ที่เชื่อมต่อระหว่างเว็บเพจในเว็บกราฟ G
ε^*	=	เซตของลิงค์ที่เชื่อมต่อระหว่างเว็บเพจในเว็บกราฟย่อย G^*
δ	=	ค่าน้อยที่สุดที่ยอมรับได้ที่ใช้ตรวจตอบการลู่เข้าของสมการเพจแร็งค์
$\vec{A}_i$	=	เวกเตอร์ที่เป็นตัวแทนค่าออเทอริตี้ของเว็บเพจทั้งหมดในเว็บกราฟ ณ รอบการคำนวณที่ i
ACB_i	=	ค่าที่แสดงถึงคุณสมบัติการเพิ่มค่าคะแนนเพจแร็งค์ของบูซฟาร์มกลุ่มที่ i
AGL_i	=	จำนวนลิงค์เฉลี่ยที่ชี้มายังเว็บเพจของเว็บเพจที่เป็นการสแปมระบบสืบค้นข้อมูลในช่วงที่ i จากเว็บเพจที่ไม่เป็นการสแปมระบบสืบค้นข้อมูล
APR_i	=	ค่าคะแนนเพจแร็งค์เฉลี่ยของเว็บเพจที่ไม่เป็นการสแปมระบบสืบค้นข้อมูลแต่มีลิงค์ชี้มายังเว็บเพจที่เป็นการสแปมระบบสืบค้นข้อมูลในช่วงที่ i
B_v	=	เซตของเว็บเพจที่ชี้ไปยังเว็บเพจ v
$B(BF_i)$	=	เซตของเว็บเพจที่มีลิงค์ชี้มายังเว็บเพจที่เป็นสมาชิกของบูซฟาร์มกลุ่มที่ i
BF	=	เซตของเว็บเพจที่เป็นสมาชิกของบูซฟาร์ม
BF_i	=	เซตของเว็บเพจในบูซฟาร์มกลุ่มที่ i
BFM_i	=	เมตริกซ์บูซฟาร์มของบูซฟาร์มกลุ่มที่ i
c	=	ค่า damping factor มีค่าอยู่ระหว่าง [0,1]
C	=	ค่าที่ใช้ในการนอมอลไลท์

คำอธิบายตัวแปรที่ใช้ในงานวิจัย (ต่อ)

$\vec{d}$	=	เวกเตอร์ที่เป็นตัวแทนเว็บเพจที่มีจำนวนลิงค์ชี้ออกเท่ากับ 0
$damping(t)$	=	ฟังค์ชั่นลดค่า
$\vec{E}$	=	เทเลพอตเวกเตอร์ซึ่งเป็นตัวแทนการสุ่มเดินไปยังเว็บเพจใดๆ ในเว็บกราฟ
$\vec{E}^*$	=	เทเลพอตเวกเตอร์ที่เบี่ยงค่าสุ่มเดินไปยังเว็บเพจที่ดี
G	=	เว็บกราฟที่มีทิศทาง
G^*	=	เว็บกราฟย่อยที่มีทิศทาง
G_p	=	เซตของเว็บเพจที่ไม่เป็นการสแปมโครงสร้างลิงค์ที่มีลิงค์ชี้มายังเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์ p
$\vec{H}_i$	=	เวกเตอร์ที่เป็นตัวแทนค่าฮับของเว็บเพจทั้งหมดในเว็บกราฟ ณ รอบการคำนวณที่ i
k	=	ขนาดของบิตเวกเตอร์ ยกตัวอย่างเช่น 32 บิต
K_i	=	รอบการคำนวณเพจแร็งค์ที่ลู่เข้าของเว็บกราฟจำลองของบูซฟาร์มกลุ่มที่ i
$l(t)$	=	จำนวนลิงค์ที่ใช้ในการสุ่มเดิน t
L_p	=	เซตของลิงค์จากเว็บเพจที่ดีที่ชี้มายังเว็บเพจ p
M	=	เมตริกซ์ความสัมพันธ์ที่เป็นตัวแทนโครงสร้างลิงค์เว็บกราฟที่กำหนดให้เว็บเพจที่มีจำนวนลิงค์ชี้ออกเท่ากับ 0 มีลิงค์ชี้ออกไปยังทุกๆเว็บเพจในเว็บกราฟ และเมื่อพิจารณาหลัก (column) แสดงด้วยเว็บเพจต้นทาง และแถว (row) แสดงด้วยเว็บเพจปลายทาง
M_e	=	เมตริกซ์ความสัมพันธ์ที่เป็นตัวแทนโครงสร้างลิงค์เว็บกราฟที่พิจารณา
M^*	=	เมตริกความสัมพันธ์ที่เป็นตัวแทนเว็บกราฟที่ทำการกลับลิงค์
M_b	=	เมตริกซ์ความสัมพันธ์ที่ปรับลดผลกระทบของบูซฟาร์ม
non_BFM	=	เมตริกซ์ที่ไม่มีบูซฟาร์ม
N	=	จำนวนเว็บเพจทั้งหมดในเว็บกราฟที่พิจารณา

คำอธิบายตัวแปรที่ใช้ในงานวิจัย (ต่อ)

N_i	=	จำนวนเว็บเพจที่เป็นการสแปมโครงสร้างลิงค์ในช่วงการกระจายตัวของเว็บเพจ i
O(p)	=	ฟังค์ชั่นที่เป็นตัวแทนการตรวจสอบเว็บเพจโดยผู้เชี่ยวชาญ
P	=	เมตริกซ์ความสัมพันธ์ที่เป็นตัวแทนโครงสร้างลิงค์เว็บกราฟที่กำหนดให้เว็บเพจที่มีจำนวนลิงค์ชี้ออกเท่ากับ 0 มีลิงค์ชี้ออกไปยังทุกๆเว็บเพจในเว็บกราฟ และเมื่อพิจารณาหลัก (column) แสดงด้วยเว็บเพจปลายทาง และแถว (row) แสดงด้วยเว็บเพจต้นทาง
$Penalty$	=	เทเลพอตเวกเตอร์ที่เบี่ยงค่าสุ่มเดินไปยังเว็บเพจที่มีการกระจายตัวของเว็บเพจช่วยเหลือแตกต่างจากการกระจายตัวแบบพาวเวอร์ลอร์
$PPR_q(p)$	=	การคำนวณค่าคะแนนเพจแร็งค์ของเว็บเพจ p ใดๆที่เบี่ยงเบนค่าสุ่มกระโดดในเทเลพอตเวกเตอร์ ณ ตำแหน่ง q^{th} มีค่าเท่ากับ 1 นอกจากนั้นมีค่าเป็น 0 ทั้งหมด
$P[t]$	=	ฟังค์ชั่นการสุ่มเดินในเว็บกราฟ
$PageRank(j_n)$	=	ค่าคะแนนเพจแร็งค์ของเว็บเพจ j ณ รอบการคำนวณที่ n
$Pr(i_n)$	=	ผลรวมค่าคะแนนเพจแร็งค์ของเว็บเพจที่เป็นสมาชิกของบูซฟาร์มกลุ่มที่ i ในรอบการคำนวณที่ n
$\text{Rank}_i(u)$	=	ค่าคะแนนเพจแร็งค์ของเว็บเพจ u
$\vec{R}_i$	=	เวกเตอร์ที่เป็นตัวแทนค่าเพจแร็งค์ของเว็บเพจทั้งหมดในเว็บกราฟ ณ รอบการคำนวณที่ i
$Randomjump(i_n)$	=	ค่าสุ่มเดินจากเว็บเพจ x ไปยังทุกเว็บเพจที่เป็นสมาชิกของบูซฟาร์มกลุ่มที่ i ในรอบการคำนวณที่ n
$\vec{S}_i$	=	เวกเตอร์ที่เป็นตัวแทนค่าสแปมแร็งค์ของเว็บเพจทั้งหมดในเว็บกราฟ ณ รอบการคำนวณที่ i
t	=	การสุ่มเดินตามลิงค์ในเว็บกราฟ
T_{io}	=	ค่าน้อยที่สุดที่ยอมรับได้ที่ใช้ตรวจสอบการเป็นสมาชิกของลิงค์ฟาร์ม

คำอธิบายตัวแปรที่ใช้ในงานวิจัย (ต่อ)

T_{pp}	=	ค่าน้อยที่สุดที่ยอมรับได้ที่ใช้ตรวจสอบการเป็นสมาชิกของเว็บเพจช่วยเหลือของลิงค์ฟาร์ม
V	=	เซตของเว็บเพจในเว็บกราฟ G
V^*	=	เซตของเว็บเพจในเว็บกราฟย่อย G^*
V_x	=	บิตเวกเตอร์ของเว็บเพจ x
VM_i	=	เมตริกซ์ลิงค์เสมือน
$\vec{W}$	=	ฟังค์ชั่นการจัดลำดับ

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นายกำธร พันธุมะผล
วัน เดือน ปี ที่เกิด	วันที่ 16 มิถุนายน 2525
สถานที่เกิด	กรุงเทพมหานคร
ประวัติการศึกษา	วศ.บ. (วิศวกรรมคอมพิวเตอร์) คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (พ.ศ.2546)
ตำแหน่งหน้าที่การงานปัจจุบัน	
สถานที่ทำงานปัจจุบัน	
ผลงานดีเด่นและรางวัลทางวิชาการ	
ทุนการศึกษาที่ได้รับ	