ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

ปริญญา

วิทยาการคอมพิวเตอร์ วิทยาการคอมพิวเตอร์

สาขา **ภาควิชา**

เรื่อง อัลกอริทึมการตรวจจับเว็บอนาจารด้วยลิงก์ฟาร์ม

Link Farm Based Pornographic Web Detection Algorithm

นามผู้วิจัย พันเอก สุเอซ อมาตยกุล

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์สุขุมาล กิติสิน, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

(ผู้ช่วยศาสตราจารย์ชวลิต ศรีสถาพรพัฒน์, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

(รองศาสตราจารย์สุรศักดิ์ สงวนพงษ์, วศ.ม.)

หัวหน้าภาควิชา

(ผู้ช่วยศาสตราจารย์ศิริกร จันทร์นวล, M.S.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ ..... **เดือน** ..... **พ.ศ.** .....

วิทยานิพนธ์

เรื่อง

อัลกอริทึมการตรวจจับเว็บอนาจารด้วยลิงก์ฟาร์ม

Link Farm Based Pornographic Web Detection Algorithm

โดย

พันเอก สุเอซ อมาตยกุล

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
พ.ศ. 2552

สุเอซ อมาตยกุล, พันเอก 2552: อัลกอริทึมการตรวจจับเว็บอนาจารด้วยลิงก์ฟาร์ม ปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์) สาขาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: ผู้ช่วยศาสตราจารย์ สุขุมาล กิติสิน, Ph.D. 62 หน้า

การศึกษาวิจัยนี้มีวัตถุประสงค์เพื่อหาแนวทางลดภาระของผู้เชี่ยวชาญในการรวบรวมรายชื่อเว็บไซต์อนาจาร โดยได้นำเสนออัลกอริทึมสำหรับตรวจจับเว็บไซต์อนาจารแบบกึ่งอัตโนมัติ เรียกว่า Link Farm Based Pornographic Web Detection Algorithm (LFPD) LFPD ทำงานโดยใช้ผู้เชี่ยวชาญตรวจหาเว็บไซต์อนาจารเพื่อเป็นข้อมูลเริ่มต้น จากนั้นทำงานต่อไปโดยการตรวจสอบโครงสร้างเส้นเชื่อมระหว่างเว็บไซต์ เพื่อหาเว็บไซต์ที่มีเส้นเชื่อมชี้ไปและกลับร่วมกับกลุ่มของเว็บไซต์อนาจารที่ตรวจพบก่อนหน้า ซึ่งสามารถแบ่งการทำงานของ LFPD ได้เป็นสามขั้น ได้แก่ การกำหนดกลุ่มเว็บไซต์อนาจารเริ่มต้นโดยผู้เชี่ยวชาญ การรวบรวมข้อมูลเว็บ และการตรวจจับเว็บไซต์อนาจาร

ผลการทดลองพบว่าจากเว็บไซต์อนาจารเริ่มต้นที่ระบุโดยผู้เชี่ยวชาญจำนวน 8 เว็บไซต์ เมื่อเราให้ LFPD ทำงานจำนวน 3 รอบ สามารถตรวจจับเว็บไซต์อนาจารได้ตั้งแต่ 18,725 - 44,123 เว็บไซต์ ขึ้นอยู่กับค่า T ซึ่งอยู่ในช่วงตั้งแต่ 1 ถึง 4 โดยเมื่อ T มีค่ามากขึ้นจำนวนเว็บไซต์ที่ตรวจจับได้ก็จะลดลงตามลำดับ เมื่อพิจารณาขีดความสามารถของ LFPD พบว่าในรอบที่หนึ่งและสองมีผลการตรวจจับที่ความแม่นยำสูงกว่าร้อยละ 90 ในทุกๆ ค่าของ T อย่างไรก็ตามจำนวนเว็บไซต์อนาจารสูงสุดที่ตรวจจับได้ด้วยความแม่นยำสูงกว่าร้อยละ 90 เป็นผลการทำงานในรอบที่สามเมื่อใช้ T 4 โดยตรวจจับเว็บไซต์อนาจารได้จำนวน 18,725 เว็บไซต์ นอกจากนี้เรายังพบว่ามีเว็บไซต์ทั่วไปที่มิใช่เว็บไซต์อนาจารแต่มีความสัมพันธ์ระหว่างกัน โดยมีการสร้างเส้นเชื่อมชี้ไปและกลับร่วมกับเว็บไซต์อนาจารโดยเจตนา

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Suez Amatyakul, Colonel 2009: Link Farm Based Pornographic Web Detection Algorithm. Master of Science (Computer Science), Major Field: Computer Science, Department of Computer Science. Thesis Advisor: Assistant Professor Sukumal Kitisin, Ph.D. 62 pages.

The purpose of this research is to find a better solution and more effective way to detect and identify pornographic websites. This research proposes a a semi-automatic technique called Link Farm Based Pornographic Web Detection Algorithm (LEPD) algorithm as a mean for examining and identifying pornographic websites. The algorithm will assist experts to search for a set of pornographic websites based on an initial seed set. The LEPD will then examining the structures of the websites in the seed set to find any links that will lead to connections with other pornographic websites previously identified. LEPD works in three steps; first, experts identify a group of pornographic websites as a seed set; second, the algorithm automatically collects pornographic websites' information and structures; third, the algorithm examines and identifies whether the collected URLs are pornographic websites.

The result from this research founded that from the 8 websites initially identified by expertise. When we run LEPD 3 times, we were able to detect pornographic websites from 18,725 to 44,123 websites. The number of websites detected is depended on value of T which is in the range between 1 to 4. As the value of T increases the number of websites detected decreases. When we consider the capability of LEPD it founded that in the first and second run the accuracy rate of detection was higher than 90 percent in every mean of T. However, the highest number (18,725) of pornographic websites that LFPD was able to detect with accuracy rate of 90 percent or higher was from the third run where T equal to 4. Furthermore, we founded that there are other websites that are not pornographic websites which intentionally link back and forth with pornographic websites; for example, commercial and match-making services.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ดีด้วยความช่วยเหลืออย่างดียิ่งของ ผู้ช่วยศาสตราจารย์สุขุมาล กิติสิน อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก ที่ได้สั่งสอน ให้คำแนะนำ วิธีการและขั้นตอนในการดำเนินงาน ชี้แนวทางแก้ไขปัญหาและตรวจแก้ข้อบกพร่องต่างๆ ในงานวิจัยตลอดมา ขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ชวลิต ศรีสถาพรพัฒน์ และรองศาสตราจารย์สุรศักดิ์ สงวนพงษ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่กรุณาให้คำแนะนำ และข้อเสนอแนะในการทำวิทยานิพนธ์นี้

ขอกราบขอบพระคุณ คุณพ่อสมชาย คุณแม่ละเอียด อมาตยกุล ผู้ให้กำเนิดและเลี้ยงดูลูกมาเป็นอย่างดีโดยตลอด ขอขอบคุณทุกคนในครอบครัวโดยเฉพาะอย่างยิ่ง คุณวจิราพร อมาตยกุล ที่ได้เป็นกำลังใจและให้การสนับสนุนในการศึกษาต่อระดับปริญญาโท และขอขอบพระคุณคณาจารย์ทุกท่านที่ให้การอบรมสั่งสอนวิชาความรู้ต่างๆ แก่ศิษย์

ท้ายที่สุดขอขอบคุณเพื่อนรุ่นที่ 6 รุ่นพี่ทุกคน ผู้บังคับบัญชาและเพื่อนร่วมงานที่คอยช่วยเหลือและให้คำแนะนำอย่างดีมาโดยตลอดห้วงการศึกษา ทำให้วิทยานิพนธ์นี้สำเร็จลุล่วงไปได้ด้วยดี และขอขอบคุณโครงการปริญญาโทภาคพิเศษที่ช่วยอำนวยความสะดวกในการประสานงานทั้งด้านอุปกรณ์และสถานที่ตลอดระยะเวลาการศึกษา

สุเอซ อมาตยกุล

มีนาคม 2552

สารบัญ

หน้า

สารบัญตาราง

สารบัญภาพ

อัลกอริทึมการตรวจจับเว็บอนาจารด้วยลิงก์ฟาร์ม

Link Farm Based Pornographic Web Detection Algorithm

คำนำ

ปัจจุบันการใช้งานอินเทอร์เน็ตโดยเฉพาะบริการประเภทเวิลด์ไวด์เว็บได้ขยายตัวอย่างกว้างขวาง ทำให้เราสามารถสื่อสารข้อมูลอันเป็นประโยชน์ต่อประชาชนกลุ่มต่างๆ ได้อย่างสะดวก แต่ในทางกลับกันก็เป็นช่องทางการเผยแพร่ข้อมูลที่ไม่เหมาะสม โดยเฉพาะอย่างยิ่งการเผยแพร่ข้อมูลที่มีลักษณะอนาจารต่อกลุ่มเยาวชน ซึ่งที่ผ่านมาได้มีการป้องกันการเข้าถึงเว็บไซต์อนาจารโดยใช้การตรวจหาและคัดกรองเว็บเพจด้วยการตรวจจับข้อความและภาพอนาจาร รวมทั้งการป้องกันโดยใช้บัญชีรายชื่อเว็บไซต์อนาจารจากฐานข้อมูลที่มีผู้ให้บริการ หรือรายชื่อเว็บไซต์ที่ผู้ใช้กำหนด อย่างไรก็ตามปัญหาสำคัญในการจัดทำบัญชีรายชื่อเว็บไซต์อนาจารได้แก่ การรวบรวมและตรวจสอบรายชื่อเว็บไซต์อนาจารซึ่งดำเนินการโดยใช้ผู้เชี่ยวชาญทำให้สิ้นเปลืองแรงงานและเวลาเป็นอย่างมาก

เพื่อลดภาระในการตรวจหาเว็บไซต์อนาจาร เราจึงได้ศึกษาโครงสร้างของเว็บไซต์อนาจาร และพบว่า เว็บไซต์ในกลุ่มดังกล่าวมีการสร้างเส้นเชื่อมระหว่างกลุ่มเว็บไซต์ประเภทเดียวกันเป็นจำนวนมาก ทั้งนี้เพื่อประโยชน์ในการนำทางผู้ใช้ไปสู่เว็บไซต์ภายในกลุ่ม และยังเป็นการเพิ่มอันดับที่ปรากฏบนเครื่องมือค้นหาข้อมูลบนเว็บ (search engine) ให้กับเว็บไซต์ของตน รูปแบบของเส้นเชื่อมระหว่างเว็บไซต์ดังกล่าวทำให้กลุ่มของเว็บไซต์อนาจารมีโครงสร้างคล้ายกับลิงก์ฟาร์ม (link farm) ที่เกิดจากการสแปมหรือสแปมฟาร์ม ซึ่งมีผู้ศึกษาเกี่ยวกับการตรวจหาสแปมฟาร์ม โดยการค้นหาเว็บเพจที่ถูกสแปมหรือสแปมเพจจากเว็บเพจที่มีเส้นเชื่อมชี้ไปและกลับระหว่างกัน (reciprocal link) ซึ่งในงานวิจัยดังกล่าวเรียกว่า common link และเรียกเว็บเพจทั้งคู่ว่า common node เมื่อเว็บเพจที่พิจารณามี common node สูงกว่าค่าที่กำหนดถือเว็บเพจดังกล่าวเป็นสแปมเพจ จากนั้นจึงนำข้อมูลสแปมเพจที่ได้ไปใช้ในการคำนวณเพื่อปรับการจัดอันดับเว็บเพจให้ถูกต้องต่อไป

ดังนั้นในการศึกษาวิจัยนี้จะนำเทคนิคการตรวจจับสแปมเพจด้วยการตรวจสอบโครงสร้างเส้นเชื่อมข้างต้นมาปรับใช้ เพื่อตรวจจับเว็บไซต์อนาจารในลักษณะกึ่งอัตโนมัติ ซึ่งจะสามารถลด

ระยะเวลาและแรงงานในการตรวจหากลุ่มของเว็บไซต์อนาจารได้เป็นอย่างมาก เมื่อเปรียบเทียบกับการตรวจสอบด้วยผู้เชี่ยวชาญเพียงอย่างเดียว โดยการทำงานในขั้นแรกเราจะใช้ผู้เชี่ยวชาญตรวจหากลุ่มของเว็บไซต์อนาจารมาจำนวนหนึ่ง เพื่อใช้เป็นกลุ่มเว็บไซต์อนาจารเริ่มต้น (seed set) จากนั้นจะรวบรวมข้อมูลเว็บจากเส้นเชื่อมที่ชี้ออกจาก seed set และตรวจหาเว็บไซต์ที่มีเส้นเชื่อมชี้ไปและกลับ (common link) ร่วมกับเว็บไซต์อนาจารที่เป็น seed set หากเว็บไซต์ใดมี common link จำนวนมากกว่าขีดต่ำสุดที่กำหนด ถือว่าเว็บไซต์ดังกล่าวเป็นเว็บไซต์อนาจาร และเพิ่มรายชื่อเว็บไซต์ดังกล่าวเข้าสู่ seed set จากนั้นจะวนรอบการทำงานต่อไป จนกว่าจะครบวงรอบที่กำหนดหรือไม่พบเว็บไซต์อนาจารเพิ่มเติม

วัตถุประสงค์

วิทยานิพนธ์นี้มีวัตถุประสงค์เพื่อศึกษาและออกแบบอัลกอริทึมในการตรวจจับเว็บไซต์อนาจารที่ทำงานแบบกึ่งอัตโนมัติ โดยใช้ผู้เชี่ยวชาญร่วมกับการตรวจสอบความสัมพันธ์ของเว็บไซต์ด้วยโครงสร้างเส้นเชื่อม เพื่อลดระยะเวลาและแรงงานในการตรวจหาเว็บไซต์อนาจารให้น้อยลง เมื่อเปรียบเทียบกับการตรวจหาเว็บไซต์อนาจารโดยใช้ผู้เชี่ยวชาญเพียงอย่างเดียว ทั้งนี้อัลกอริทึมที่พัฒนาขึ้นและข้อมูลเว็บไซต์อนาจารที่รวบรวมได้สามารถนำไปใช้ประโยชน์ในการพัฒนาระบบป้องกันการเข้าถึงเว็บไซต์ที่ไม่เหมาะสมได้ต่อไป

การตรวจเอกสาร

1. ปัญหาและการควบคุมการเผยแพร่เว็บไซต์อนาจาร

อินเทอร์เน็ตเป็นเครือข่ายของเครือข่ายที่เชื่อมต่อคอมพิวเตอร์จากทั่วทุกมุมโลก แต่ละจุดเชื่อมต่อสามารถติดต่อกันผ่านเส้นทางต่างๆ ที่ไม่แน่นอนตายตัว ข้ามเขตแดนและอำนาจทางกฎหมายของประเทศต่างๆ เราจึงไม่สามารถควบคุมหรือกำกับดูแลอินเทอร์เน็ตในลักษณะเดียวกันกับสื่อแบบเก่า และจากจำนวนผู้ใช้งานอินเทอร์เน็ตทั่วโลกที่เพิ่มสูงขึ้นกว่าหนึ่งพันล้านคน (Internet World Stats, 2008) ทำให้อินเทอร์เน็ตเป็นสื่อที่มีอิทธิพลอย่างสำคัญต่อประชาคมโลก

จากคุณลักษณะของอินเทอร์เน็ตที่เนื้อหาที่เผยแพร่ออกสู่ผู้ใช้อาจไม่ได้รับการควบคุมและกลั่นกรองอย่างเพียงพอเมื่อเทียบกับสื่อประเภทอื่น โดยเฉพาะการเผยแพร่เนื้อหาลามกอนาจาร ดังนั้นการเปิดรับข้อมูลจากอินเทอร์เน็ตของเด็กและเยาวชน จึงเป็นประเด็นที่ต้องให้ความสนใจ โดยปัจจุบันมีกลไกที่ใช้ในการกำกับดูแลเนื้อหาอินเทอร์เน็ต ประกอบด้วย กฎหมาย มาตรการบทลงโทษ (Legal sanction), การปิดกั้นและกลั่นกรองเนื้อหา (Blocking and filtering system), กฎ กติกา มารยาท (Codes of conduct), สายด่วน (Hotlines) และการรู้เท่าทันสื่อ (Media literacy) (พิรงรอง และนิธิมา, 2547)

ในประเทศไทยการกำกับดูแลอินเทอร์เน็ตอย่างเป็นทางการเป็นหน้าที่ของกระทรวงเทคโนโลยีสารสนเทศและการสื่อสาร (ICT) โดยเข้ามารับผิดชอบต่อจากศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ที่เคยดูแลนโยบายด้านอินเทอร์เน็ตก่อนหน้าที่จะมีการจัดตั้งกระทรวงเทคโนโลยีสารสนเทศและการสื่อสารขึ้นมาในปี พ.ศ. 2545 และเมื่อประมาณกลางปี พ.ศ. 2546 กระทรวงเทคโนโลยีสารสนเทศและการสื่อสารได้มีการจัดตั้งหน่วยงานกำกับดูแลใหม่ ใช้ชื่อว่าคณะกรรมการสืบสวน ป้องกัน และปราบปรามอาชญากรรมคอมพิวเตอร์ หรือ Cyber-inspector ซึ่งทำหน้าที่ตรวจสอบดูแลเนื้อหาที่มีความเสี่ยงต่างๆ บนอินเทอร์เน็ต รวมไปถึงอาชญากรรมทางอินเทอร์เน็ตต่างๆ ทั้งนี้ Cyber-inspector จะเป็นผู้ตัดสินใจว่าจะปิดกั้นเว็บไซต์อะไรบ้าง สมาชิกของ Cyber-inspector เป็นเจ้าหน้าที่ระดับสูงของหน่วยงานราชการ และองค์กรธุรกิจที่มีความสนใจและมีทักษะในการใช้อินเทอร์เน็ต (พิรงรอง และนิธิมา, 2547)

นอกจากนี้ กระทรวงเทคโนโลยีสารสนเทศและการสื่อสารยังมีสายด่วนสำหรับรับเรื่องร้องเรียนเกี่ยวกับเนื้อหาหรือพฤติกรรมที่ผิดกฎหมายหรือเป็นอันตรายบนอินเทอร์เน็ต โดยประชาชนทั่วไปสามารถเข้าถึงสายด่วนดังกล่าวได้ทั้งทางโทรศัพท์และทางเว็บไซต์ ปัจจุบัน กระทรวงเทคโนโลยีสารสนเทศและการสื่อสารได้รวบรวมเว็บไซต์ที่มีเนื้อหาเป็นอันตรายอยู่ในบัญชีดำ (มีการปิดกั้นแล้ว) ประมาณ 1,300 เว็บไซต์ จากจำนวนประมาณ10,000 เว็บไซต์ที่ได้มีการแจ้งเข้ามา (พิรงรอง และนิธิมา, 2547)

2. หลักการที่เกี่ยวข้อง

ในส่วนของหลักการที่เกี่ยวข้องประกอบด้วย ระบบการกรองเว็บ โครงสร้างเว็บ และการสแปมเว็บ

2.1 ระบบการกรองเว็บที่ใช้ในปัจจุบันแบ่งได้เป็น 4 ลักษณะใหญ่ คือ ระบบการกรองเว็บโดยใช้บัญชีรายชื่อเว็บไซต์ (URL blocking) ระบบการกรองเว็บโดยใช้การตรวจสอบคำสำคัญ (Keyword blocking) ระบบการกรองเว็บโดยใช้การตรวจสอบภาพ (Image blocking) และระบบการกรองเว็บโดยใช้การจัดอันดับเว็บ (Rating system) ซึ่งแต่ละระบบมีรายละเอียดดังนี้

2.1.1 ระบบการกรองเว็บโดยใช้บัญชีรายชื่อเว็บไซต์ (URL blocking) ระบบการกรองเว็บนี้เป็นการกรองโดยอาศัยบัญชีรายชื่อเว็บที่ควรถูกปิดกั้นหรือบัญชีรายชื่อเว็บที่ไม่ปลอดภัย (Blacklists) ซึ่งมีการจัดทำและเผยแพร่ เช่น บัญชีรายชื่อเว็บที่จัดทำโดย URLBlacklist.com (URLBlacklist.com, 2008) หรืออาจใช้รายการเว็บที่อนุญาตให้เข้าไปใช้งานได้หรือบัญชีรายชื่อเว็บที่ปลอดภัย (Whitelists) แต่ระบบการกรองเว็บรูปแบบนี้มีปัญหาที่สำคัญสองประการ ประการแรก เว็บไซต์ที่ไม่ปลอดภัยเหล่านี้มีเป็นจำนวนมากและเพิ่มจำนวนขึ้นเรื่อยๆ ดังนั้นจึงต้องยอมสูญเสียประสิทธิภาพการทำงานของระบบคอมพิวเตอร์ในการเข้าชมเว็บไปกับการตรวจสอบรายชื่อเว็บในบัญชี ประการที่สอง ข้อมูลเว็บไซต์เหล่านี้มีการเปลี่ยนแปลงทุกวัน ดังนั้นจึงจำเป็นต้องค้นหาและรวบรวมข้อมูลรายชื่อเว็บไซต์ที่ไม่ปลอดภัยเพิ่มเติมอย่างต่อเนื่อง ทำให้การสร้างระบบกรองเว็บรูปแบบนี้มีค่าใช้จ่ายสูงและจำเป็นต้องใช้เวลาและแรงงานจำนวนมาก เนื่องจากส่วนใหญ่การรวบรวมรายชื่อเว็บไซต์เหล่านี้มักใช้การตรวจสอบด้วยผู้เชี่ยวชาญ (Lee et al., 2002)

2.1.2 ระบบการกรองเว็บโดยใช้การตรวจสอบคำสำคัญ (Keyword blocking) ระบบการกรองเว็บรูปแบบนี้เป็นการใช้รายการคำสำคัญ (Keyword) ที่จัดทำขึ้นเพื่อระบุเว็บเพจที่ไม่เหมาะสม เมื่อตรวจพบคำหรือข้อความดังกล่าว เว็บเพจนั้นจะถูกป้องกันมิให้เข้าถึง จากการศึกษาของ Lee และคณะ (Lee et al., 2002) พบว่าเว็บเพจที่มีเนื้อหาอนาจารมักปรากฏคำหรือข้อความลักษณะอนาจารที่ตำแหน่งต่างๆ ในเว็บเพจ ได้แก่ ชื่อเว็บเพจ (Web page title), ข้อความเตือนการเข้าถึงเนื้อหาภายในเว็บไซต์ (Warning message block), ข้อความที่ปรากฏให้เห็นบนเว็บเพจ, ข้อมูลที่บรรจุในคุณสมบัติ description และ keywords ขององค์ประกอบ meta, URL และส่วนขยาย (embedded) ของ URL และข้อความที่ปรากฏในคุณสมบัติ alt ขององค์ประกอบ img

อย่างไรก็ตามระบบการกรองเว็บด้วยการใช้คำสำคัญมีข้อจำกัดสามประการที่สำคัญ ประการแรก ได้แก่ ปัญหาเกี่ยวกับความหมายของคำ (meaning of words) ที่อาจก่อให้เกิดการปิดกั้นเกินความจำเป็น (over blocking) ตัวอย่างเช่น กรณีการใช้คำว่า “sex” เป็นคำสำคัญจะทำให้เว็บที่เกี่ยวกับเพศศึกษา (sex education) อาจถูกปิดกั้น เนื่องจากปรากฏคำว่า “sex” ในเนื้อหาของเว็บไซต์ดังกล่าว ซึ่งแท้ที่จริงแล้วเว็บไซต์มิได้มุ่งหมายที่จะแสดงเนื้อหาในลักษณะอนาจารแต่อย่างใด ปัญหาที่สอง ได้แก่ ข้อผิดพลาดในการสะกดคำ ซึ่งอาจเกิดจากการที่ผู้พัฒนาเว็บไซต์ต้องการหลีกเลี่ยงระบบการกรองเว็บ จึงสะกดคำต่างๆ ที่ต้องการให้เป็นคำสำคัญอย่างผิดพลาด ทำให้ตรวจจับเว็บไซต์นั้นไม่ได้ เช่น เว็บอนาจารบางเว็บจะแทนคำว่า “pornorgraphic” ด้วยคำว่า “pornorgaphic” เพื่อให้เกิดความสับสนในเรื่องของคำ และทำให้เกิดอุปสรรคต่อระบบการกรองเว็บโดยใช้การตรวจสอบคำสำคัญ (Lee et al., 2002) ปัญหาสุดท้าย หากเว็บไซต์อนาจารถูกจัดทำขึ้นโดยมิได้ใช้ภาษาที่กำหนดไว้ในรายการคำสำคัญ เช่น ภาษาอังกฤษ หรือภาษาไทย แต่ใช้ภาษาในท้องถิ่นของผู้ผลิตเว็บไซต์ ก็จะทำให้ไม่สามารถกรองเว็บดังกล่าวด้วยระบบนี้ได้

2.1.3 ระบบการกรองเว็บโดยใช้การตรวจสอบภาพ (Image blocking) เนื่องจากเนื้อหาที่เผยแพร่ทางเว็บไซต์อนาจารส่วนมากเป็นการนำเสนอภาพนิ่งหรือภาพเคลื่อนไหวเกี่ยวกับร่างกายของมนุษย์ โดยภาพดังกล่าวมักแสดงให้เห็นผิวหนังส่วนใหญ่ของร่างกาย ดังนั้นเราจึงตรวจจับเว็บเพจอนาจารได้โดยการตรวจจับภาพที่พื้นที่ส่วนใหญ่ปรากฏเป็นสีผิวมนุษย์ (skin detection) (Chan et al., 1999) (Lin et al., 2003) อย่างไรก็ตามวิธีการดังกล่าวอาจมีปัญหาอุปสรรคในการตรวจจับภาพชุดว่ายน้ำหรือภาพในลักษณะเดียวกันที่ปรากฏผิวหนังมนุษย์เป็นส่วนมาก แต่แท้จริงแล้วมิได้เป็นภาพลามกอนาจาร

2.1.4 ระบบการกรองเว็บโดยใช้การจัดอันดับเว็บ (Rating system) เป็นการกลั่นกรองเว็บโดยให้ผู้พัฒนาเว็บหรือองค์กรอิสระที่เกี่ยวข้องเป็นผู้พิจารณาเนื้อหาของเว็บ โดยติดสัญลักษณ์ (label) ที่ระบุประเภทของเนื้อหาดังกล่าวไว้ที่เว็บเพจของเว็บไซต์ เมื่อผู้ใช้เข้าชมเว็บดังกล่าว ซอฟต์แวร์ที่เกี่ยวข้องจะตรวจสอบสัญลักษณ์กับค่าที่ผู้ใช้กำหนด เพื่อพิจารณาว่าเว็บเพจดังกล่าวสมควรที่จะถูกนำเสนอต่อผู้ชมหรือไม่

PICS (Platform for Internet Content Selection) (Resnick and Miller, 2002) เป็นระบบการทำสัญลักษณ์ (label) ให้แก่เว็บ เพื่อแสดงคุณสมบัติของข้อมูลที่ปรากฏบนเว็บนั้นๆ ลักษณะของการทำสัญลักษณ์มี 2 รูปแบบ แบบแรก ได้แก่ Self-Rating เป็นการที่เจ้าของเว็บเป็นผู้จัดอันดับหรือประเภทเว็บของตน โดยอาศัยข้อมูลหรือสารสนเทศของเจ้าของเว็บเป็นองค์ประกอบหลักในการพิจารณา วิธีถัดมาคือ Third-Party Rating เป็นการให้องค์กรอิสระเข้ามาเป็นผู้กำหนดหรือประมาณค่าของเว็บ สัญลักษณ์ที่กำหนดจะปรากฏในลักษณะของ MIME type การกลั่นกรองเว็บกระทำโดยใช้ซอฟต์แวร์เพื่อการคัดเลือก (selection software) ซึ่งเป็นซอฟต์แวร์แยกต่างหากจากระบบ PICS ทั้งนี้ซอฟต์แวร์เพื่อการคัดเลือกมีหน้าที่ในการตรวจสอบสัญลักษณ์ของเว็บที่จะเข้าชมว่าเป็นประเภทที่อนุญาตหรือไม่อนุญาตให้เข้าชม แต่ปัญหาของระบบกรองเว็บในลักษณะนี้ได้แก่ กรณีที่การจัดอันดับเว็บผิดพลาดหรือไม่เหมาะสม จะทำให้ระบบการกรองเว็บนั้นไม่น่าเชื่อถือและขาดความแม่นยำ

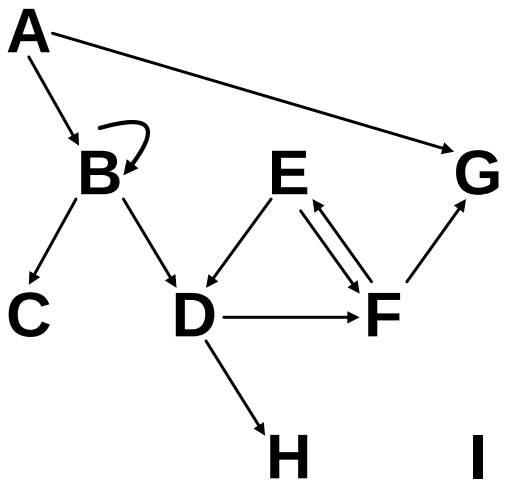
Internet Content Rating Association หรือ ICRA (www.icra.org) เป็นองค์กรอิสระที่ไม่แสวงหากำไรมีสำนักงานอยู่ในสหรัฐอเมริกาและอังกฤษ ภารกิจหลักของ ICRA คือ การปกป้องเด็กจากสิ่งอันตรายไปพร้อมกับการปกป้องเสรีภาพในการแสดงออกของผู้ผลิตเนื้อหาในอินเทอร์เน็ต ระบบการกลั่นกรองเนื้อหาทางอินเทอร์เน็ตของ ICRA มีทั้งระบบกรองเนื้อหา การป้องกันเว็บไซต์ที่มีรายชื่อตามบัญชีที่กำหนด และการจัดอันดับและติดสัญลักษณ์ (Self-Classification and labeling) ระบบการจัดอันดับและติดสัญลักษณ์ของ ICRA มีรากฐานอยู่บนการจัดแบ่งประเภทด้วยตนเองของผู้ให้บริการอินเทอร์เน็ตโดยใช้คำศัพท์ (vocabulary) ของ ICRA ซึ่งจะพยายามอธิบายเนื้อหาอย่างไร้อัตวิสัย (objectivity) ทั้งนี้ ICRA จะออกแบบสอบถามให้ผู้ให้บริการเนื้อหากรอกและประเมินจากค่าคะแนนที่ได้ในแบบสอบถามว่า เนื้อหาในเว็บไซต์นั้นควรถูกอธิบายหรือติดป้าย (label) ว่าอย่างไร เช่น sexual material (เนื้อหาเกี่ยวกับเรื่องเพศ) violence (ความรุนแรง) chat (การพูดคุย) การติดป้ายบอกเนื้อหาของผู้ให้บริการสามารถเปรียบได้กับการที่ผู้ผลิตอาหารติดป้ายบอกส่วนผสมของอาหารบนบรรจุภัณฑ์ของอาหารนั้น

ทั้งนี้ป้ายอิเล็กทรอนิกส์ (electronic label) ของ ICRA สามารถอ่านได้โดยเว็บบราวเซอร์ทั่วไป ปัจจุบันมีเว็บไซต์มากกว่า 90,000 เว็บไซต์ทั่วโลกที่ลงทะเบียนใช้ระบบการจัดประเภทและติดป้ายเนื้อหาของ ICRA (Internet Content Rating Association, 2008) (พิรงรอง และนิธิมา, 2547)

2.2 โครงสร้างเว็บ (web structure)

โครงสร้างเว็บเปรียบได้กับกราฟแบบมีทิศทาง (direct graph) เราอาจเทียบเว็บโหนด ซึ่งหมายถึง เว็บเพจ เว็บไดเร็คทอรี่ หรือเว็บไซต์ อย่างใดอย่างหนึ่งขึ้นอยู่กับระดับที่พิจารณา กับโหนด (node) ของกราฟ และไฮเปอร์ลิงก์ (hyperlink) หรือลิงก์ (link) หรือเส้นเชื่อมระหว่างเว็บโหนด เทียบกับเส้นเชื่อม (egde) ของกราฟ และนำหลักการของกราฟมาอธิบายความสัมพันธ์ระหว่างเว็บโหนดและเส้นเชื่อมได้ดังนี้ (Björneborn and Ingwersen, 2004)

1. inlink หมายถึง เส้นเชื่อมที่ชี้เข้ามายังเว็บโหนดที่พิจารณา
2. outlink หมายถึง เส้นเชื่อมที่ชี้ออกจากเว็บโหนดที่พิจารณาไปยังเว็บโหนดอื่น
3. self-link หมายถึง เส้นเชื่อมที่ชี้ออกจากเว็บโหนดที่พิจารณา และชี้กลับมายังเว็บโหนดเดิม
4. isolated หมายถึง เว็บโหนดที่โดดเดี่ยวโดยไม่มีทั้งเส้นเชื่อมชี้เข้าและออก
5. reciprocal link หมายถึง เว็บโหนดสองเว็บโหนดที่มีเส้นเชื่อมระหว่างกันอย่างน้อยสองเส้น ซึ่งชี้ในทิศทางตรงข้ามกัน
6. triadically interlinked หมายถึง เว็บโหนดสามเว็บโหนดขึ้นไปที่มีเส้นเชื่อมชี้เข้าและออกจากเว็บโหนดในกลุ่มเป็นวงกลม
7. transversal link หรือ shortcut หมายถึง เป็นเส้นเชื่อมที่เชื่อมระหว่างเว็บโหนดสองเว็บโหนดที่อยู่ต่างกลุ่มกัน ในลักษณะเส้นทางลัด (shortcut)
8. reachable หมายถึง เว็บโหนดที่มีเส้นเชื่อมเชื่อมต่อกันโดยผ่านเว็บโหนดอื่น
9. colink หมายถึง เว็บโหนดสองเว็บโหนดที่มีเส้นเชื่อมชี้มาจากเว็บโหนดอันหนึ่ง หรือเว็บโหนดสองเว็บโหนดที่มีเส้นเชื่อมชี้ไปยังเว็บโหนดอันเดียวกัน



ภาพที่ 1 โครงสร้างเว็บประกอบด้วยเว็บโหนดและเส้นเชื่อม

จากภาพที่ 1 แสดงเว็บโหนดจำนวน 9 เว็บโหนด แต่ละเว็บโหนดมีความสัมพันธ์กัน โดย B ได้รับเส้นเชื่อมชี้เข้าหรือ inlink จาก A, B มีเส้นเชื่อมชี้ออกหรือ outlink ไปยัง C, B มีเส้นเชื่อมชี้กลับมายังตนเองหรือ self-link, A ไม่มีเส้นเชื่อมชี้เข้าหรือ inlink, C ไม่มีเส้นเชื่อมชี้ออกหรือ outlink, I ไม่มีทั้งเส้นเชื่อมชี้เข้าและชี้ออกจึงเป็นเว็บโหนดที่โดดเดี่ยว, E และ F มีเส้นเชื่อมชี้ระหว่างกันในทิศทางตรงกันข้ามหรือ reciprocal link, D E และ F ทั้งสามเว็บโหนดมีเส้นเชื่อมชี้เชื่อมต่อกันเป็นวงกลมหรือ triadically interlinked, A มีเส้นเชื่อมไปยัง G ซึ่งเป็นเว็บโหนดที่อยู่ต่างกลุ่มกันในลักษณะของ transversal link หรือ shortcut, H เป็นเว็บโหนดที่เชื่อมต่อกันกับ A โดยผ่านเว็บโหนดอื่นหรือ reachable, C และ D ถูกชี้โดยเส้นเชื่อมจาก B หรือ colinked โดย B, B และ E มีเส้นเชื่อมชี้ไปยัง D หรือ colinking ต่อ D

2.3 การสแปมเว็บ (web spam) การสแปมเว็บเป็นความพยายามที่จะทำให้เครื่องมือค้นหาข้อมูลบนเว็บ (search engine) ได้รับข้อมูลนอกเหนือจากที่ควร เพื่อทำให้เว็บเพจเป้าหมายได้รับการจัดอันดับสูงกว่าปกติ เว็บเพจที่ถูกสร้างหรือถูกปรับปรุงข้อมูลโดยการสแปมเรียกว่าสแปมเพจ รูปแบบการสแปมที่กระทำต่อเครื่องมือค้นข้อมูลบนเว็บมีการดำเนินการแยกเป็นสองเทคนิค ได้แก่ Boosting Techniques และ Hiding Techniques (Gyongyi and Garcia-Molina, 2005)

2.3.1 Boosting Techniques เป็นเทคนิคการสร้างสแปมเพจ โดยปรับปรุงเว็บเพจที่มีอยู่เดิม หรือสร้างสแปมเพจขึ้นมาใหม่ เพื่อทำให้สแปมเพจหรือเว็บเพจเป้าหมายได้รับการจัดอันดับ

จากเครื่องมือค้นหาข้อมูลบนเว็บสูงยิ่งขึ้น เทคนิคดังกล่าวประกอบด้วย การสแปมเนื้อหาเว็บ และการสแปมโครงสร้างเว็บ

การสแปมเนื้อหาเว็บเป็นเทคนิคการปรับปรุงเนื้อหาของเอกสาร เพื่อทำให้สแปมเพจมีข้อมูลตรงกับคำค้นที่ต้องการ เทคนิคนี้ใช้ได้ผลกับเครื่องมือค้นหาข้อมูลบนเว็บที่คำนวณและจัดอันดับเว็บเพจโดยให้น้ำหนักคะแนนตามคำค้นและตำแหน่งของคำค้นที่ปรากฏ การสแปมเนื้อหาเว็บเพจแบ่งตามตำแหน่งของเนื้อหาที่ทำการสแปมได้ดังนี้

1. Body spam เป็นการแทรกคำสำคัญ (keyword) เข้าไปในบริเวณตัว (body) ของเว็บเพจ

2. Title spam เป็นการแทรกคำสำคัญเข้าไปในส่วนชื่อ (title) ของเว็บเพจ

3. Meta tag spam เป็นการใส่คำสำคัญเข้าไปในในคุณสมบัติ keyword หรือ description ขององค์ประกอบ meta ของเว็บเพจ

4. Anchor text spam เป็นการใส่คำหรือข้อความสำคัญไว้ในคำอธิบายเส้นเชื่อม (link label)

5. URL spam เป็นการใส่คำหรือข้อความสำคัญไว้ใน URL ทั้งนี้ URL ดังกล่าวมักมีความยาวมากกว่าปกติ

การสแปมเนื้อหาเว็บนอกจากแบ่งตามตำแหน่งของเนื้อหาที่ทำการสแปมแล้ว ยังสามารถแบ่งตามลักษณะการดำเนินการได้เป็น 4 ประเภท ดังนี้

1. Repetition เป็นการทำซ้ำคำสำคัญบางคำ โดยจะปรากฏคำดังกล่าวจำนวนมากในเว็บเพจ เพื่อเพิ่มผลในการจัดอันดับของเครื่องมือค้นหาข้อมูลบนเว็บ เทคนิคนี้จะได้ผลจำกัดเฉพาะคำค้นที่ตรงกับคำที่มีการทำซ้ำ

2. Dumping เป็นการนำคำสำคัญจำนวนมาก ซึ่งไม่จำเป็นต้องมีความสัมพันธ์หรือเกี่ยวข้องกับเนื้อหาของเอกสารมาบรรจุในเว็บเพจ คำดังกล่าวอาจนำมาจากพจนานุกรมหรือแหล่งข้อมูลอื่น เพื่อให้มีอย่างน้อยหนึ่งคำที่ตรงกับคำค้นจากเครื่องมือค้นหาข้อมูลบนเว็บ เทคนิคนี้จะใช้ได้กับคำค้นใดๆ แต่อาจไม่ส่งผลในการเพิ่มค่าในการจัดอันดับได้มากนัก

3. Weaving เป็นการสอดแทรกคำสำคัญแบบสุ่มไว้ในเนื้อหาต้นฉบับของเว็บเพจ ทั้งนี้หากไม่สอดแทรกคำจนมากเกินไป และสแปมเมอร์ทำการกระจายคำสำคัญอย่างเหมาะสม เทคนิคนี้จะสามารถป้องกันการตรวจจับการทำซ้ำคำสำคัญจากเครื่องมือค้นหาข้อมูลบนเว็บได้

4. Phrase stitching เป็นการนำประโยคหรือวลีที่มีเนื้อหาสำคัญจากเอกสารต่างๆ ซึ่งสแปมเมอร์คาดว่าจะได้รับความสนใจจากเครื่องมือค้นหาข้อมูลบนเว็บ มาเชื่อมต่อกันโดยไม่ต้องคำนึงถึงความหมายโดยรวม สแปมเมอร์จะใช้เทคนิคนี้เมื่อต้องการสร้างสแปมเพจอย่างรวดเร็ว

การสแปมโครงสร้างเว็บเป็นการปรับปรุงโครงสร้างเว็บ เพื่อเพิ่มเส้นเชื่อมให้กับสแปมเพจหรือเว็บเพจเป้าหมาย เทคนิคนี้ใช้ได้ผลกับเครื่องมือค้นหาข้อมูลบนเว็บที่คำนวณและจัดอันดับโดยให้น้ำหนักคะแนนตามจำนวนเส้นเชื่อมที่ชี้เข้าหรือออกจากเว็บเพจ ทั้งนี้เว็บเพจเป้าหมายที่ถูกใช้ในการสแปมแบ่งได้เป็นสามลักษณะ ดังนี้

1. Inaccessible pages หมายถึง เว็บเพจที่สแปมเมอร์ไม่สามารถหรือไม่มีสิทธิในการปรับแก้ข้อมูล ซึ่งทำให้ไม่สามารถปรับปรุงหรือเปลี่ยนแปลง outlink ของเว็บเพจดังกล่าวได้

2. Accessible pages หมายถึง เว็บเพจที่ตามปกติอยู่ภายใต้การควบคุมของผู้อื่น เช่น เว็บมาสเตอร์ แต่ผู้ควบคุมเปิดโอกาสให้สแปมเมอร์สามารถเพิ่มเติมหรือแก้ไขข้อมูลได้อย่างจำกัด เช่น การโพสต์ข้อความเข้าสู่เว็บเพจ เป็นต้น ข้อความที่สแปมเมอร์โพสต์นั้นจะบรรจุเส้นเชื่อมที่เชื่อมไปยังสแปมเพจ ดังนั้นสแปมเมอร์จึงสามารถสร้าง outlink ได้จากเว็บเพจดังกล่าว

3. Own pages หมายถึง เว็บเพจที่สแปมเมอร์ควบคุมดูแลเอง สแปมเมอร์จึงสามารถที่จะควบคุมเนื้อหาและการสร้างเส้นเชื่อมสำหรับเว็บเพจในกลุ่มนี้ได้อย่างเต็มที่ ซึ่งเราเรียกเว็บเพจในกลุ่มนี้ว่า สแปมฟาร์ม (spam farm)

การสแปมโครงสร้างเว็บโดยใช้ outlink นั้น สแปมเมอร์จะพยายามสร้าง outlink จำนวนมากจากสแปมเพจให้ชี้ไปยังเว็บเพจที่เป็นที่นิยมหรือรู้จักกันดี (good authority) ซึ่งจะเป็นการเพิ่ม hub score ให้กับสแปมเพจดังกล่าว สำหรับการสแปมโครงสร้างเว็บโดยใช้ inlink นั้นเป็นการพยายามสร้าง inlink จำนวนมากไปยังเว็บเพจหรือกลุ่มของเว็บเพจที่เป็นเป้าหมายอันใดอันหนึ่ง เทคนิคในการสแปมโครงสร้างเว็บโดยใช้ inlink มีหลายวิธี เช่น

1. Honey pot เป็นการสร้างเว็บเพจหรือกลุ่มของเว็บเพจที่บรรจุข้อมูลที่เป็นที่ต้องการของผู้ใช้ เช่น คู่มือการใช้งานโปรแกรมที่สำคัญ เป็นต้น โดยสแปมเมอร์จะซ่อนเส้นเชื่อมที่ชี้ไปยังสแปมเพจหรือกลุ่มของสแปมเพจที่เป็นเป้าหมายไว้ในเว็บเพจดังกล่าว

2. Infiltrate a web directory เว็บไซต์บางแห่งจะอนุญาตให้เพิ่มเส้นเชื่อมเข้าสู่เว็บไซต์ภายใต้หัวข้อหรือไดเร็คทอรีที่กำหนด ดังนั้นสแปมเมอร์ จึงอาจแทรกเส้นเชื่อมที่ชี้ไปยังเว็บเพจเป้าหมายได้

3. Comment สแปมเมอร์อาจรวม URL ไว้เป็นส่วนหนึ่งของข้อความที่โพสต์เข้าสู่ blogs, boards, guest books หรือ wikis

4. Link exchange เป็นการแลกเปลี่ยนเส้นเชื่อมระหว่างกลุ่มของสแปมเมอร์เพื่อให้มีเส้นเชื่อมเข้ามายังเว็บที่ทำการสแปมเพิ่มมากขึ้น

5. Expired domain เมื่อโดเมนเนมหมดอายุ แต่เส้นเชื่อมจากเว็บไซต์ต่างๆ ที่เชื่อมมายัง URL ภายในโดเมนข้างต้นจะยังคงอยู่จนกว่าผู้ดูแลเว็บอื่นๆ จะตรวจพบและดำเนินการแก้ไขหรือลบเส้นเชื่อมดังกล่าว ดังนั้นสแปมเมอร์จึงจะซื้อโดเมนที่หมดอายุแล้วมาใช้ และอาจสร้างสแปมเพจเพิ่มเติมขึ้นอีก ซึ่งจะทำให้เว็บไซต์ดังกล่าวได้เปรียบกลุ่มของสแปมเพจตามปกติ เนื่องจากมีเส้นเชื่อมเดิมจากเว็บไซต์ต่างๆ ชี้เข้ามาแล้วจำนวนหนึ่ง

6. Spam farm เมื่อสแปมเมอร์มีความสามารถในการสร้างเว็บเพจภายในเว็บไซต์หรือสามารถควบคุมเว็บไซต์ สแปมเมอร์จะสร้างเส้นเชื่อมจำนวนมากระหว่างกลุ่มของเว็บเพจในเว็บไซต์ดังกล่าว ซึ่งจะเป็นการเพิ่มอันดับของเว็บเพจภายในกลุ่ม เมื่อค้นหาด้วยเครื่องมือค้นหาข้อมูลบนเว็บที่ใช้การจัดอันดับด้วยโครงสร้างเส้นเชื่อม

2.3.2 Hiding Techniques เป็นเทคนิคในการซ่อนการกระทำของสแปมเมอร์จากการพบเห็นของผู้ใช้และผู้ดูแลเครื่องมือค้นหาข้อมูลบนเว็บที่พยายามตรวจหาการสแปมดังกล่าว การทำ Hiding Techniques มีหลายลักษณะด้วยกัน ได้แก่

Content Hiding เป็นการทำให้สีของคำหรือข้อความหรือเส้นเชื่อมที่ใช้ในการทำสแปมมีความกลมกลืนกับพื้นหลังของเว็บเพจ โดยอาจใช้วิธีการสร้างจุดเชื่อมด้วยภาพขนาดเล็ก เช่น 1 x 1 pixel หรือการกำหนดให้ภาพมีลักษณะโปร่งใส เป็นต้น

Cloaking เป็นการที่เว็บเซิฟเวอร์ของสแปมเมอร์จะทำการตรวจสอบคำขอที่ได้รับ หากคำขอนั้นมาจากผู้ใช้ตามปกติ เว็บเซิฟเวอร์ก็จะส่งข้อมูลเว็บเพจตามปกติไปให้ แต่ถ้า

คำขอดังกล่าวมาจาก spider ของเครื่องมือค้นหาข้อมูลบนเว็บ เว็บเซิฟเวอร์ของสแปมเมอร์ก็จะส่งข้อมูลเว็บเพจที่บรรจุเนื้อหาที่ทำการสแปมไว้ไปให้

Redirection เป็นการซ่อนเนื้อหาที่ใช้ในการสแปมไว้ในเว็บเพจที่สแปมเมอร์กำหนดให้บราวเซอร์ทำการ redirect โดยอัตโนมัติไปยัง URL อื่นทันทีที่เว็บเพจนั้นถูกโหลด ซึ่งจะทำให้ผู้ใช้ไม่ทันเห็นข้อมูลดังกล่าว แต่เครื่องมือค้นหาข้อมูลบนเว็บจะตรวจพบและนำเว็บเพจนั้นไปทำการจัดอันดับ

3. งานวิจัยที่เกี่ยวข้อง

ในส่วนของงานวิจัยที่เกี่ยวข้องประกอบด้วย การศึกษาเกี่ยวกับโครงสร้างเส้นเชื่อมของเว็บไซต์อนาจาร และการศึกษาการตรวจจับสแปมเพจซึ่งประกอบด้วย อัลกอริทึม TrustRank โดย Gyongyi และคณะ และการตรวจจับสแปมเพจภายในลิงก์ฟาร์มโดย Wu และ Davison

3.1 การศึกษาเกี่ยวกับโครงสร้างเส้นเชื่อมของเว็บไซต์อนาจาร

ในการศึกษาวิจัยเกี่ยวกับการเชื่อมโยงระหว่างเว็บไซต์ Krishna Bharat และคณะได้กล่าวถึงสแปม (spam) ไว้ว่า “การสแปมโครงสร้างเว็บเป็นการสร้างเส้นเชื่อมเพื่อเพิ่มผลการจัดอันดับโดยเครื่องมือค้นหาข้อมูลบนเว็บให้กับเว็บไซต์ที่ต้องการ ซึ่งการสแปมดังกล่าวมักปรากฏโดยเฉพาะในกลุ่มของเว็บไซต์อนาจาร (pornographic web sites)” (Bharat et al., 2001)

ในปี 1998 (Page et al., 1998) กลุ่มนักวิจัยนำโดย Larry Page และคณะได้ศึกษาเกี่ยวกับการคำนวณหาความสำคัญของเว็บเพจ โดยมีแนวความคิดว่าถ้าเว็บเพจต้นทางใดชี้ไปยังเว็บเพจปลายทางใดแล้วนั้น หมายถึงผู้สร้างเว็บเพจต้นทางได้ส่งมอบความสำคัญให้กับเว็บเพจปลายทางนั้น วิธีการดังกล่าวมีชื่อเรียกว่า PageRank หลักการสำคัญของ PageRank มีสองประการ ประการแรก ถ้าเว็บเพจใดถูกอ้างอิงหรือมีเส้นเชื่อมมาจากเว็บเพจอื่นจำนวนมาก จะถือว่ามีค่าความสำคัญมาก ประการที่สอง เว็บเพจใดที่ถูกอ้างอิงจากเว็บเพจที่มีค่าความสำคัญมากก็จะมีค่าความสำคัญมากด้วยเช่นกัน ทั้งนี้การคำนวณค่าความสำคัญจากการวิเคราะห์เส้นเชื่อมจะพิจารณาโครงสร้างของเส้นเชื่อม โดยไม่จำเป็นต้องพิจารณาเนื้อหาของเว็บเพจแต่อย่างใด

ในการศึกษาดังกล่าวข้างต้น Page และคณะได้เปรียบเทียบผลการจัดอันดับโดย PageRank กับข้อมูลการใช้งานเว็บ (web usage) พบว่าในกลุ่มของเว็บไซต์อนาจารผลลัพธ์ที่ได้ไม่สอดคล้องกัน เมื่อตรวจสอบต่อไปจึงพบว่ามีข้อมูลการใช้งานเว็บไซต์อนาจารเป็นจำนวนมาก แต่ค่า PageRank ของกลุ่มเว็บไซต์อนาจารกลับคำนวณได้ไม่สูงนัก จากผลการทดลองดังกล่าว Page และคณะจึงสรุปว่า “ในขณะที่คนทั่วไปมักเข้าใช้เว็บไซต์อนาจาร แต่ก็ไม่ต้องการที่จะสร้างเส้นเชื่อมที่ชี้จากเว็บไซต์ของเขาไปยังเว็บไซต์อนาจาร”

3.2 TrustRank

กลุ่มของนักวิจัยนำโดย Gyongyi และคณะ (Gyongyi et al., 2004) นำเสนออัลกอริทึมชื่อว่า TrustRank ซึ่งเป็นเทคนิคการแยกเว็บเพจที่ดีจากสแปมเพจแบบกึ่งอัตโนมัติ โดยมีสมมติฐานว่าเว็บเพจที่ดีมักชี้ไปยังเว็บเพจที่ดีด้วยกัน และน้อยครั้งที่จะชี้ไปยังเว็บเพจที่ไม่ดี การคำนวณ TrustRank นั้น เริ่มต้นโดยให้ผู้เชี่ยวชาญตรวจสอบหาเว็บเพจที่ดี(good pages) โดยการทำ transition matrix กลับทิศทางการชี้ของแต่ละเว็บเพจแล้วคำนวณหาค่า inverse PageRank จากนั้นคำนวณหา bias PageRank ตามจำนวนรอบที่ต้องการ โดยสร้างเว็คเตอร์ของข้อมูลที่ได้จากลำดับค่า inverse PageRank รวมกับข้อมูลเว็บเพจที่ดีที่ผู้เชี่ยวชาญระบุ ทำการปรับเว็คเตอร์เข้าสู่บรรทัดฐาน (normalize) ที่ค่าเท่ากับ 1 ผลลัพธ์ที่ได้จะสามารถระบุได้ว่าเว็บเพจใดเป็นสแปมเพจ ในขั้นตอนสุดท้ายจะทำการแยกสแปมเพจออกจากลิงก์ฟาร์ม และคำนวณหาค่า PageRank ที่ถูกต้องต่อไป

จุดเด่นของ TrustRank อยู่ที่การเลือกเว็บเพจเริ่มต้นซึ่งกระทำโดยผู้เชี่ยวชาญทำให้มั่นใจได้ว่าเว็บเพจที่ถูกเลือกเป็นเว็บเพจดี อย่างไรก็ตามก็เป็นการยากที่จะเลือกกลุ่มของเว็บเพจดีให้เพียงพอที่จะเป็นตัวแทนของเว็บไซต์แต่ละกลุ่ม (Wu et al., 2006)

3.3 การตรวจจับสแปมเพจภายในลิงก์ฟาร์ม

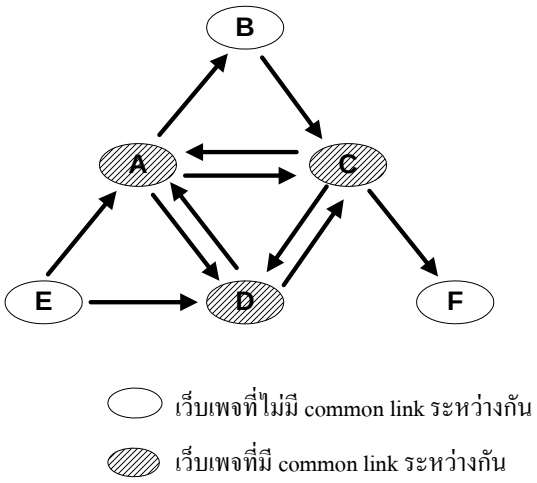
ลิงก์ฟาร์ม (link farm) หมายถึง กลุ่มของเว็บเพจที่มีเส้นเชื่อมระหว่างกันเป็นจำนวนมาก สแปมฟาร์ม (spam farm) เป็นลิงก์ฟาร์มถูกสร้างขึ้นมาเพื่อเพิ่มผลในการจัดอันดับของเว็บเพจภายในกลุ่ม เมื่อค้นหาด้วยเครื่องมือค้นหาข้อมูลบนเว็บที่จัดอันดับด้วยโครงสร้างเส้นเชื่อม ดังนั้นเพื่อให้การจัดอันดับเว็บเพจของเครื่องมือค้นหาข้อมูลบนเว็บเป็นไปอย่างเที่ยงตรง จึงมีการ

ศึกษาวิจัยเกี่ยวกับการตรวจหาลิงก์ฟาร์มที่เกิดจากการสแปม ซึ่งข้อมูลลิงก์ฟาร์มที่ได้ จะนำไปใช้ในการปรับปรุงผลการจัดอันดับของเครื่องมือค้นหาข้อมูลบนเว็บให้ถูกต้องยิ่งขึ้นต่อไป

Wu และ Davison (Wu and Davison, 2005) นำเสนออัลกอริทึมในการตรวจจับสแปมเพจภายในลิงก์ฟาร์มแบบอัตโนมัติ โดยตรวจสอบข้อมูลจากกลุ่มของเว็บเพจทั้งหมดที่รวบรวมได้ (data set) เพื่อหากลุ่มของเว็บเพจเริ่มต้น (seed set) ที่มีโครงสร้างเส้นเชื่อมประเภท reciprocal link ซึ่งในงานวิจัยนี้เรียกว่า common link แนวความคิดในการตรวจจับลิงก์ฟาร์มโดย common link มีอยู่ว่า สแปมเพจภายในลิงก์ฟาร์มจะมีเส้นเชื่อมระหว่างกันอย่างหนาแน่นและปรากฏคู่ของเว็บเพจจำนวนมากที่มีทั้ง inlink และ outlink ระหว่างกัน เรียกว่าเว็บเพจทั้งคู่ว่า common node เมื่อติดตามเส้นเชื่อมภายในลิงก์ฟาร์มที่เป็น common link ก็จะพบ common node และหากเว็บเพจที่พิจารณามี common node จำนวนมากกว่าค่าที่กำหนดก็จะถือว่าเป็นสแปมเพจและกำหนดให้เป็น seed set จากนั้นจะขยายการตรวจจับสแปมเพจภายในโครงสร้างของลิงก์ฟาร์มทั้งหมด โดยตรวจหาเว็บเพจที่มีเส้นเชื่อมชี้ออก(outlink) มายัง seed set ดังกล่าวต่อไป อัลกอริทึมดังกล่าวมีขั้นตอนการทำงานสามขั้น ได้แก่

3.3.1 การสร้าง seed set

จากแนวความคิดที่ว่าเว็บเพจภายในลิงก์ฟาร์มจะมี common link ระหว่างกัน อย่างไรก็ตามหากจะให้คู่ของเว็บเพจหรือ common node ใดๆ ที่มี common link ระหว่างกันเพียงคู่เดียวถือเป็นสแปมเพจก็จะเป็นการไม่ยุติธรรม เนื่องจากอาจมีเว็บเพจที่ดีบางเพจที่มีเส้นเชื่อมชี้ไปยังเว็บเพจที่เป็นสแปมเพจ เช่น honey pot ได้โดยไม่รู้ตัว ดังนั้นจึงกำหนดขีดเริ่มเปลี่ยน (threshold) T_{io} เพื่อใช้เป็นค่าต่ำสุดสำหรับตัดสินว่าควรจะต้องมี common node จำนวนเท่าใด เว็บเพจดังกล่าวจึงจะถือเป็นสแปมเพจและถูกกำหนดเป็น seed set ซึ่งจากการทดลองพบว่าค่า T_{io} ที่เหมาะสมเท่ากับ 3

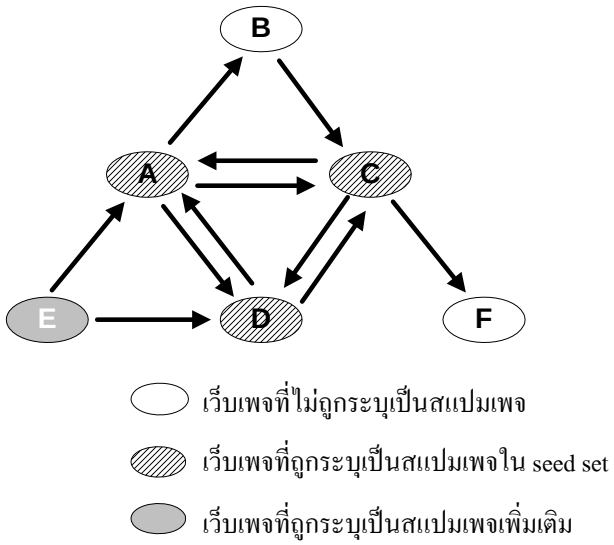


ภาพที่ 2 เว็บเพจจำนวน 6 เพจ และการตรวจหา common link

จากภาพที่ 2 แสดงให้เห็นว่าเว็บเพจ A, C และ D มี common link ระหว่างกัน เมื่อพิจารณาที่ A จะพบว่า A เป็น common node กับ C และ D ซึ่งรวมแล้ว A จะมี common node เท่ากับ 2 หาก T_{io} ที่ใช้เท่ากับ 2 ก็จะถือว่า A เป็น สแปมเพจ และนำเข้าสู่ seed set นอกจากนี้เมื่อพิจารณาต่อไปที่เว็บเพจ C และ D ก็จะพบว่า C และ D เป็นสแปมเพจในทำนองเดียวกัน

3.3.2 การขยายการตรวจจับ (Expansion step)

จาก seed set ที่ได้ การที่จะขยายการตรวจจับออกไปนั้นจำเป็นที่จะต้องค้นหา bad page เพิ่มเติมจาก data set ทั้งหมด โดยมีแนวความคิดว่า ถ้าเว็บเพจหนึ่งมีเส้นเชื่อมเส้นหนึ่งชี้ออก (outlink) ไปยังสแปมเพจแล้ว เราจะยังไม่ด่วนตัดสินว่าเว็บเพจดังกล่าวเป็นสแปมเพจ แต่ถ้าเว็บเพจหนึ่งมีเส้นเชื่อม (outlink) ชี้ไปยังสแปมเพจหลายๆ เพจ หรือเป็นการ colinked ต่อ seed set แล้ว เชื่อได้ว่าเว็บเพจดังกล่าวเป็นสแปมเพจและเป็นส่วนหนึ่งของลิงก์ฟาร์ม ดังนั้นจึงกำหนดขีดเริ่มเปลี่ยน T_{pp} (Threshold ParentPenalty) เพื่อใช้เป็นตัวกำหนดค่าต่ำสุดที่ใช้ในการตัดสินว่าจะต้องมีเส้นเชื่อมชี้ออก (outlink) ไปยังสแปมเพจเท่าใด จึงจะถูกตัดสินว่าเป็นสแปมเพจซึ่งเป็นส่วนหนึ่งของลิงก์ฟาร์ม



ภาพที่ 3 เว็บเพจจำนวน 6 เพจ และการขยายการตรวจจับ

จากภาพที่ 3 หากเรากำหนด T_{pp} เท่ากับ 2 และผลการทำงานในขั้นการสร้าง seed set พบว่าเว็บเพจ A, C และ D เป็นสแปมเพจและถูกกำหนดให้เป็น seed set แล้ว เมื่อพิจารณาที่เว็บเพจ E ก็จะพบว่า E มีเส้นเชื่อมชี้ไปยัง seed set จำนวน 2 เว็บเพจ ได้แก่ A และ D ดังนั้น E จึงเป็นสแปมเพจ สำหรับเว็บเพจ B ไม่เป็นสแปมเพจเนื่องจากมีเส้นเชื่อมชี้ไปยัง seed set เพียง 1 เส้น และเว็บเพจ F ก็ไม่เป็นสแปมเพจเช่นกันเนื่องจากไม่มีเส้นเชื่อมชี้ไปยัง seed set

3.3.3 การปรับการจัดอันดับ

เมื่อค้นพบสแปมเพจภายในกลุ่มของเว็บเพจทั้งหมด (data set) แล้ว จะนำข้อมูลที่ได้ไปใช้ในการปรับปรุงการจัดลำดับ (ranking) ซึ่งการปรับการจัดลำดับของเว็บเพจดังกล่าวนั้นอาจกระทำได้สองหนทางด้วยกัน หนทางแรกเป็นการปรับโทษเว็บเพจเหล่านี้อย่างเข้มงวดโดยการนำออกจากเว็บกราฟ แต่ในความเป็นจริงการดำเนินการดังกล่าวอาจรุนแรงเกินไป ตัวอย่างเช่น บริษัทอาจมีผลิตภัณฑ์หลายชนิด และเผยแพร่แต่ละผลิตภัณฑ์ในเว็บไซต์ที่แตกต่างกัน แต่เนื่องจากเว็บไซต์ต่างๆ ที่เผยแพร่ผลิตภัณฑ์แต่ละชนิดเป็นเว็บไซต์ของบริษัทเดียวกัน แต่ละเว็บไซต์จึงมีเส้นเชื่อมระหว่างกันคล้ายลิงก์ฟาร์ม ดังนั้นแทนที่จะปรับโทษเว็บเหล่านี้อย่างรุนแรงก็อาจดำเนินการในหนทางที่สอง ได้แก่ การปรับลดค่าน้ำหนัก (down-weight)

อุปกรณ์และวิธีการ

อุปกรณ์

1. เครื่องคอมพิวเตอร์ตั้งโต๊ะ หน่วยประมวลผลความเร็ว 2.3 กิกะเฮิร์ต, หน่วยความจำหลัก 1 กิกะไบต์, หน่วยความจำสำรอง 320 กิกะไบต์ ความเร็วรอบ 7200 รอบต่อนาที
2. โปรแกรม NetBeans IDE 6.1
3. โปรแกรม Java Development Kit (JDK) version 1.6.0_06
4. โปรแกรมจัดการฐานข้อมูล MySQL Server 5.0
5. ระบบปฏิบัติ Microsoft Windows XP Professional SP2

วิธีการ

แผนการดำเนินงานของงานวิจัยนี้ ได้แก่ การศึกษาโครงสร้างของเว็บไซต์อนาจารเพื่อใช้เป็นแนวทางในการออกแบบอัลกอริทึมสำหรับตรวจจับเว็บไซต์อนาจาร การพัฒนาโปรแกรมให้ทำงานตามอัลกอริทึมดังกล่าว จากนั้นทำการทดลองเพื่อทดสอบผลการทำงานของอัลกอริทึมโดยการรวบรวมข้อมูลเว็บจากระบบอินเทอร์เน็ต และเปรียบเทียบผลที่ได้รับภายใต้เงื่อนไขต่างๆ ซึ่งมีรายละเอียดของข้อมูลดังต่อไปนี้

1. การศึกษาเบื้องต้นเกี่ยวกับโครงสร้างของเว็บไซต์อนาจาร

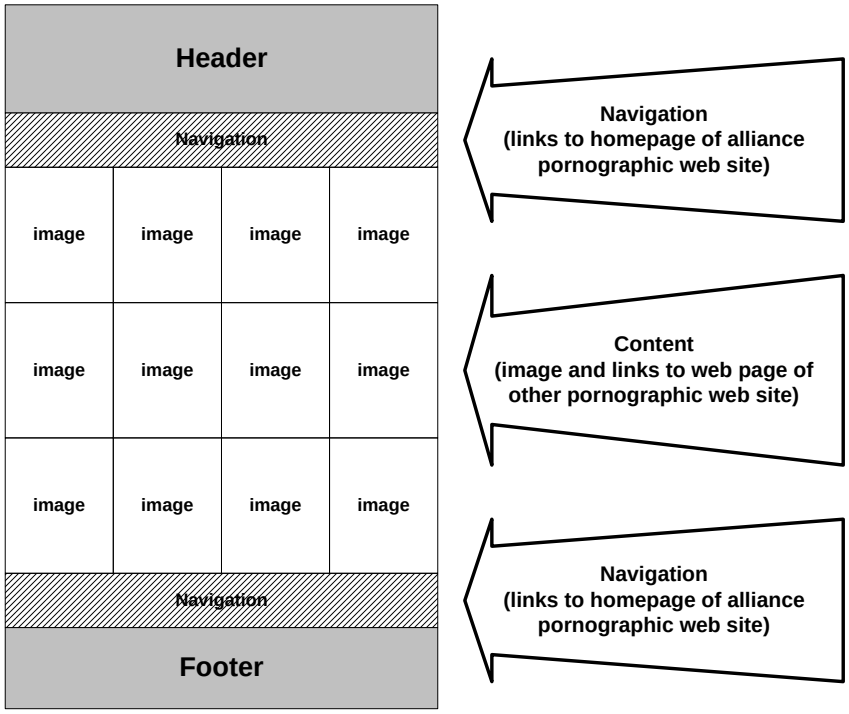
การศึกษาเบื้องต้นเกี่ยวกับโครงสร้างเว็บไซต์อนาจารเป็นการศึกษาโดยการทบทวนผลการศึกษาวิจัยที่ผ่านมาและการสังเกตเพิ่มเติมโดยผู้วิจัย เพื่อให้ทราบถึงคุณลักษณะและข้อแตกต่างระหว่างเว็บไซต์อนาจารกับเว็บไซต์ทั่วไป ซึ่งจะใช้เป็นข้อมูลสำหรับออกแบบอัลกอริทึมการตรวจจับเว็บไซต์อนาจารต่อไป

จากการทบทวนผลการศึกษาวิจัยที่ผ่านมาพบว่า ในการศึกษาวิจัยเกี่ยวกับอัลกอริทึม PageRank ของ Page และคณะ (Page et al., 1998) มีข้อสรุปว่า “ในขณะที่คนทั่วไปมักเข้าใช้เว็บไซต์อนาจาร แต่ก็ไม่ต้องการที่จะสร้างเส้นเชื่อมที่ชี้จากเว็บไซต์ของเขาไปยังเว็บไซต์อนาจาร” และในการศึกษาเกี่ยวกับการเชื่อมโยงระหว่างเว็บไซต์ Krishna Bharat และคณะ(Bharat et al.,

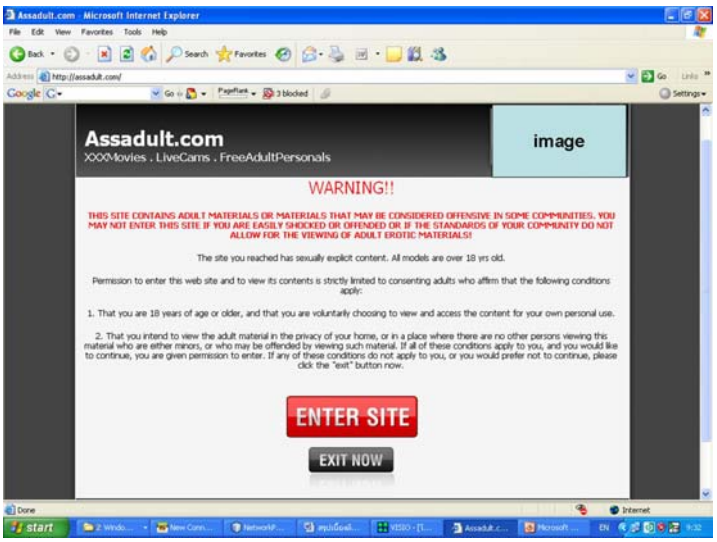
2001) ได้กล่าวถึงการสแปม (spam) ไว้ว่า “การสแปมโครงสร้างเว็บเป็นการสร้างเส้นเชื่อมเพื่อเพิ่มผลการจัดอันดับโดยเครื่องมือค้นหาข้อมูลบนเว็บให้กับเว็บไซต์ที่ต้องการ ซึ่งการสแปมดังกล่าวมักปรากฏโดยเฉพาะในกลุ่มของเว็บไซต์อนาจาร (pornographic web sites)”

เมื่อสังเกตเว็บไซต์อนาจารและเว็บไซต์ทั่วไปเพิ่มเติม เราพบลักษณะสำคัญของเว็บไซต์อนาจารสรุปได้ดังนี้

1.1 เว็บไซต์อนาจารมีการเชื่อมโยงกันเป็นกลุ่มโดยมีการสร้างเส้นเชื่อมชี้ไปยังเว็บไซต์อนาจารอื่นภายในกลุ่มเดียวกัน ทั้งนี้เส้นเชื่อมดังกล่าวมักปรากฏทั้งในโฮมเพจและเว็บเพจอื่นภายในเว็บไซต์ นอกจากนี้ยังมีเส้นเชื่อมชี้จากรูปภาพหรือข้อความเพื่อนำไปสู่ภาพหรือภาพเคลื่อนไหวที่ปรากฏในเว็บไซต์อนาจารอื่นอีกด้วย

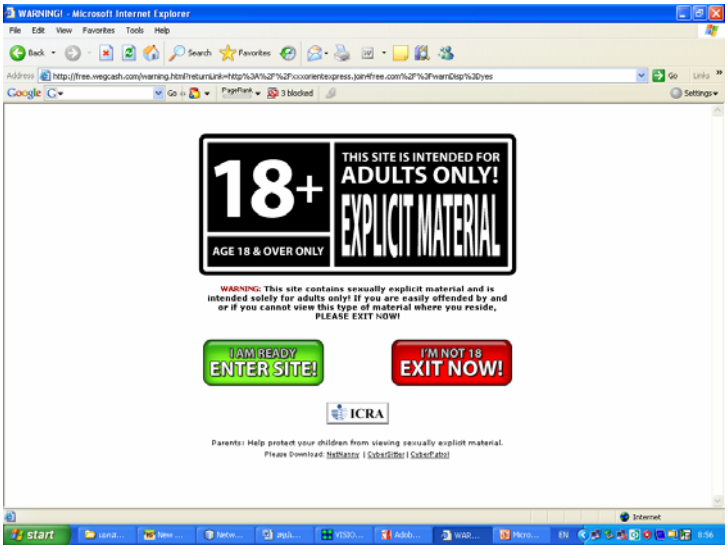


ภาพที่ 4 โครงสร้างเว็บเพจของเว็บไซต์อนาจารที่เป็นแหล่งรวบรวมเส้นเชื่อม



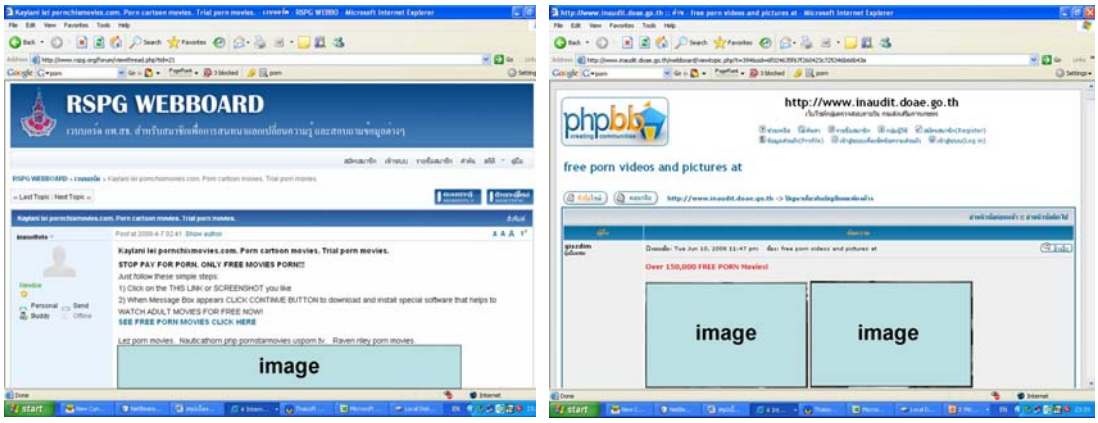
ภาพที่ 5 ตัวอย่างเว็บเพจแรกของเว็บไซต์อนาจารที่ปรากฏคำเตือนก่อนเข้าสู่เว็บไซต์

1.2 เว็บไซต์อนาจารบางเว็บไซต์จะปรากฏคำเตือนเกี่ยวกับเนื้อหาที่หน้าแรกของเว็บไซต์ เพื่อให้ผู้ใช้เลือกที่จะเข้าชมหรือไม่เข้าชมภายในเว็บไซต์ โดยข้อความดังกล่าวมักประกอบอยู่กับเส้นเชื่อมชี้เข้าไปภายในเว็บไซต์หรือชี้ออกไปยังเว็บไซต์อื่น ในเว็บเพจแรกนี้อาจไม่ปรากฏเส้นเชื่อมที่ชี้ไปยังเว็บไซต์อนาจารภายในกลุ่ม อย่างไรก็ตามเส้นเชื่อมที่ชี้ไปยังกลุ่มเว็บไซต์อนาจารมักปรากฏในเว็บเพจระดับถัดไป



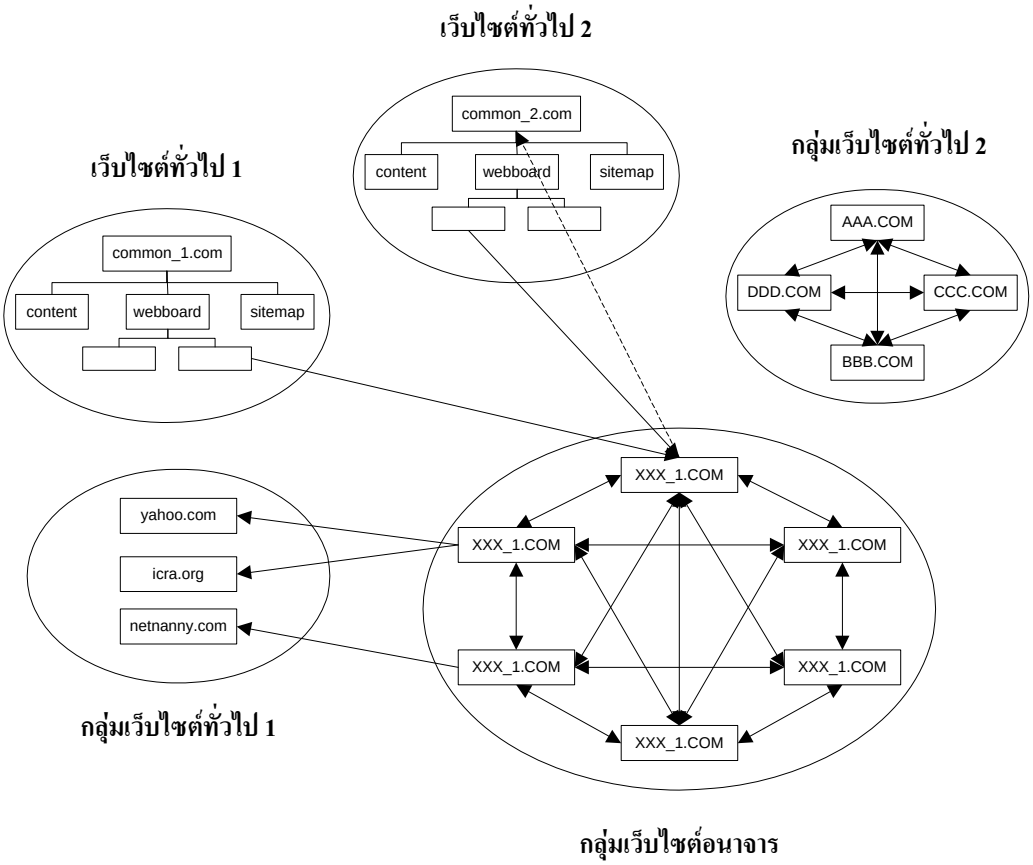
ภาพที่ 6 ตัวอย่างเว็บเพจของเว็บไซต์อนาจารที่มีเส้นเชื่อมชี้ไปยังเว็บไซต์ทั่วไป

1.3 เว็บไซต์อนาจารบางแห่งมีเส้นเชื่อมชี้ไปยังเว็บไซต์ทั่วไป เช่น google หรือ yahoo ภายใต้คำอธิบายว่าเป็นเส้นเชื่อมสำหรับผู้ที่ไม่ต้องการเข้าชมเว็บไซต์ นอกจากนี้บางแห่งยังมีเส้นเชื่อมชี้ไปยังเว็บไซต์ ICRA (www.icra.org) หรือ NETNANNY (www.netnanny.com) หรือ CYBERPATROL (www.cyberpatrol.com) ซึ่งมิใช่เว็บไซต์อนาจาร แต่เป็นเว็บไซต์ที่ให้บริการเกี่ยวกับการกรองหรือป้องกันการเข้าถึงข้อมูลเว็บที่ไม่เหมาะสม



ภาพที่ 7 เว็บเพจในเว็บไซต์ทั่วไปที่ปรากฏเส้นเชื่อมชี้ไปยังเว็บไซต์อนาจาร

1.4 เว็บไซต์ทั่วไปอาจปรากฏภาพหรือเส้นเชื่อมชี้ไปยังเว็บไซต์อนาจารได้ โดยเฉพาะอย่างยิ่งเว็บเพจเกี่ยวกับกระดานสนทนา (webboard) ของเว็บไซต์ทั่วไป



ภาพที่ 8 สมมุติฐานความสัมพันธ์ระหว่างเว็บไซต์อนาจารกับเว็บไซต์ทั่วไป

2. การตรวจจับเว็บไซต์อนาจาร

2.1 จากคุณลักษณะของเว็บไซต์อนาจารที่ปรากฏในการศึกษาวิจัยและจากการสังเกตเว็บไซต์อนาจารข้างต้น สรุปเป็นสมมุติฐานเกี่ยวกับโครงสร้างและความสัมพันธ์ระหว่างเว็บไซต์อนาจารและเว็บไซต์ทั่วไป เพื่อใช้เป็นแนวทางในการพัฒนาอัลกอริทึมของเราสามประการ ได้แก่

1. โครงสร้างความสัมพันธ์ระหว่างกลุ่มเว็บไซต์อนาจารมีลักษณะเป็นลิงก์ฟาร์ม โดยมีเส้นเชื่อมระหว่างกันอย่างหนาแน่น

2. เว็บไซต์อนาจารบางแห่งมีเส้นเชื่อมชี้ไปยังเว็บไซต์ทั่วไป

3. ปกติแล้วเว็บไซต์ทั่วไปจะไม่สร้างเส้นเชื่อมชี้กลับมายังเว็บไซต์อนาจาร

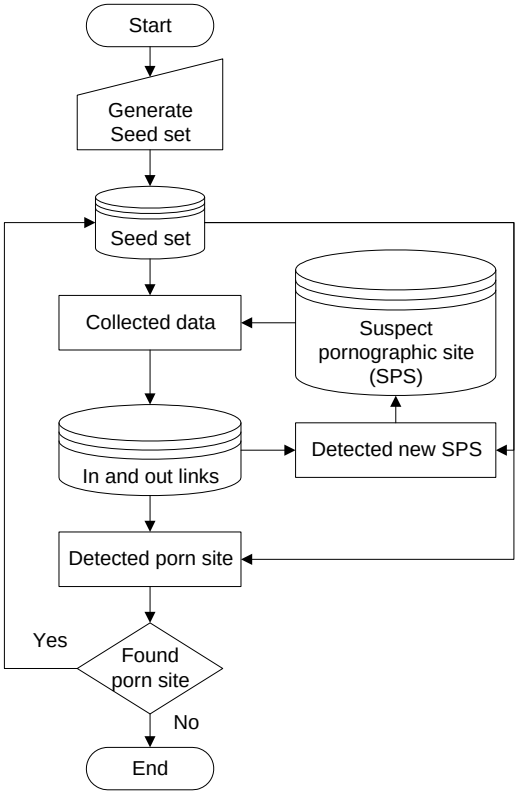
2.2 การออกแบบอัลกอริทึมการตรวจจับเว็บไซต์อนาจาร

เมื่อเราสรุปในเบื้องต้นได้ว่าโครงสร้างเส้นเชื่อมของกลุ่มเว็บไซต์อนาจารมีลักษณะคล้ายคลึงกันกับลิงก์ฟาร์ม เราจึงพัฒนาเทคนิคการตรวจจับเว็บไซต์อนาจาร เรียกว่า Link Farm Based Pornographic Web Detection Algorithm (LFPD) โดยนำการตรวจจับสแปมฟาร์มภายในลิงก์ฟาร์มมาประยุกต์ใช้ในการตรวจจับเว็บไซต์อนาจาร ด้วยการผสมผสานขั้นตอนการเลือกเว็บเพจที่ดีเพื่อเป็นข้อมูลเริ่มต้นด้วยผู้เชี่ยวชาญของ TrustRank (Gyongyi et al., 2004) และการตรวจหา seed set ของ Wu และ Davison (Wu and Davisson, 2005) เข้าด้วยกัน

การรวมขั้นตอนทั้งสองเทคนิคเป็นการนำจุดเด่นมาลบล้างจุดด้อยซึ่งกันและกัน โดยการคัดเลือกเว็บเพจที่ดีด้วยผู้เชี่ยวชาญเป็นเทคนิคที่มีมีความถูกต้องสูงกว่าการการคัดเลือกด้วยระบบอัตโนมัติ อย่างไรก็ตามข้อเสียของการใช้ผู้เชี่ยวชาญก็คือสิ้นเปลืองเวลาและแรงงาน ส่วนการตรวจหา seed set ของ Wu และ Davison เป็นเทคนิคการคัดเลือกเว็บเพจที่ไม่ดีด้วยระบบอัตโนมัติ ซึ่งมีจุดเด่นในการลดการใช้แรงงานและเวลา แต่ก็ย่อมที่จะมีความถูกต้องต่ำกว่าการคัดเลือกจากผู้เชี่ยวชาญโดยตรง ซึ่งเห็นได้จากการที่ Wu และ Davison ต้องกำหนดทางเลือกในการปรับการจัดอันดับไว้สองประการ ได้แก่ การนำเว็บเพจที่ไม่ดีออกจากเว็บกราฟ หรือการปรับลดค่าน้ำหนัก (down-weight)

จากสมมุติฐานความสัมพันธ์ระหว่างเว็บไซต์อนาจารกับเว็บไซต์ทั่วไปตามภาพที่ 8 เรามีแนวความคิดในการตรวจจับเว็บไซต์อนาจารและการคัดกรองเว็บไซต์อนาจารออกจากเว็บไซต์ทั่วไปเพื่อป้องกันการตรวจจับผิดพลาด โดยการตรวจจับเว็บไซต์อนาจารนั้นเราจะใช้ผู้เชี่ยวชาญระบุเว็บไซต์อนาจารเพื่อเป็น seed set สำหรับเริ่มต้นการทำงานของ LFPD เนื่องจากผู้เชี่ยวชาญสามารถแยกแยะกลุ่มเว็บไซต์อนาจารออกจากเว็บไซต์ทั่วไปในลักษณะของกลุ่มเว็บไซต์ทั่วไป 2 ซึ่งเป็นลิงก์ฟาร์ม ซึ่งเป็นการลดโอกาสในการกำหนด seed set ผิดพลาด ต่อจากนั้นจะตรวจจับเว็บไซต์อนาจารภายในกลุ่มโดยตรวจหาเว็บไซต์ที่มีเส้นเชื่อมชี้ไปและกลับ (reciprocal link หรือ common link) ร่วมกับเว็บไซต์อนาจารที่เป็น seed set ซึ่งจะป้องกันการตรวจจับเว็บไซต์ทั่วไปในลักษณะของเว็บไซต์ทั่วไป 1 และกลุ่มเว็บไซต์ทั่วไป 1 มิให้ถูกระบุเป็นเว็บไซต์อนาจารเนื่องจากเว็บไซต์ทั่วไป 1 และกลุ่มเว็บไซต์ทั่วไป 1 มิได้มีเส้นเชื่อมชี้ไปและกลับร่วมกับเว็บไซต์อนาจาร มีเพียงแต่เส้นเชื่อมชี้ไปหรือกลับร่วมกับเว็บไซต์อนาจารเท่านั้น สำหรับเว็บไซต์ทั่วไปที่มีลักษณะเช่นเดียวกันกับเว็บไซต์ทั่วไป 2 เราจะคัดกรองโดยการตรวจนับจำนวน

common link ซึ่งจะมีการกำหนดขีดเริ่มเปลี่ยน (Threshold) สำหรับตัดสินจำนวน common link ที่เหมาะสมในการระบุว่าเว็บไซต์ดังกล่าวเป็นเว็บไซต์อนาจารหรือไม่



ภาพที่ 9 แผนผังการทำงานของ LFPD

เราได้ออกแบบอัลกอริทึม LFPD ให้ทำงานในลักษณะเวียนเกิด (recursive) โดยแบ่งการทำงานได้เป็นสามขั้นตอน ตามภาพที่ 9 ได้แก่

1. การสร้างกลุ่มเว็บไซต์อนาจารเริ่มต้น (Generate seed set) เป็นการสร้างกลุ่มเว็บไซต์อนาจารเริ่มต้น (seed set) โดยผู้เชี่ยวชาญ

2. การรวบรวมข้อมูลเว็บ (Collected data) เป็นการรวบรวมข้อมูลเส้นเชื่อมที่ชี้เข้าและออกจาก seed set และเว็บไซต์ต้องสงสัย (suspect pornographic site, SPS)

3. การตรวจจับเว็บไซต์อนาจาร (Detected porn site) เป็นการตรวจหาเว็บไซต์อนาจารจากข้อมูลเว็บที่รวบรวมได้ ทั้งนี้หากยังคงพบเว็บไซต์อนาจารใหม่ก็จะกลับไปรวบรวมข้อมูลเว็บเพิ่มเติมจนกว่าจะครบจำนวนรอบที่กำหนดหรือจนกว่าจะไม่พบเว็บไซต์อนาจาร

```
Input

      SE    seed set expert (a set of porn site that confirm
            by expert)
      T     threshold decided porn site

Output

      S     set of porn site confirm by our algorithm + SE

Begin

(1)  S = add(SE)                          // add SE to S
(2)  C = CrawlingFromURLofSeedset (SE) // Crawling outlinks
                                             from SE
(3)
   do
      new SPS = ExtractNewDomainsToSPS (C, S)

      // new domain from OD which pointed by S is new Suspect
         pornographic site (SPS)
      SPS    = add(new SPS)              // add new SPS to SPS
      C = CrawlingFromURLofSPS (SPS)   // Crawling outlinks
                                             from SPS
      OL = outlinks(SPS, C)             // specific outlinks
                                             from SPS to SPS
      IL   = inlinks(SPS, C)            // specific inlinks
                                             from SPS to SPS

      //PS  is number of porn site that found by commonlink

      PS = commonlink(T , S, SPS, OL, IL)//detecting porn
                                           site by reciprocal
                                           links with S
   while (PS > 0)

   return S

end
```

ภาพที่ 10 อัลกอริทึม LFPD

```
function commonlink(T , S, SPS, OL, IL)

input

      T     threshold decided porn site
      S     set of porn site
      SPS   set of Suspect pornographic site
      OL    outlinks from SPS to SPS
      IL    inlinks from SPS to SPS

Output

      PS    number of domains are porn site which reciprocal
            linked with S equal to or above T

begin
      do
            m = 0
            for i = 1 to length( S ) do
                  SPS = SPS – S
                  for j = 1 to length(SPS) do
                        if (IL(SPS[j] pointing to S[i]) > 0 )
                        and(OL(S[i] pointing to SPS[j]) > 0 )
                        then
                              k j++
                                    if (  k j >= T )
                                    //this domain is porn site
                                          S add(SPS[j] )
                                          m++
                                          PS = PS + m
      while (m > 0) //if found new porn site continue :
                        else break
      return PS

end
```

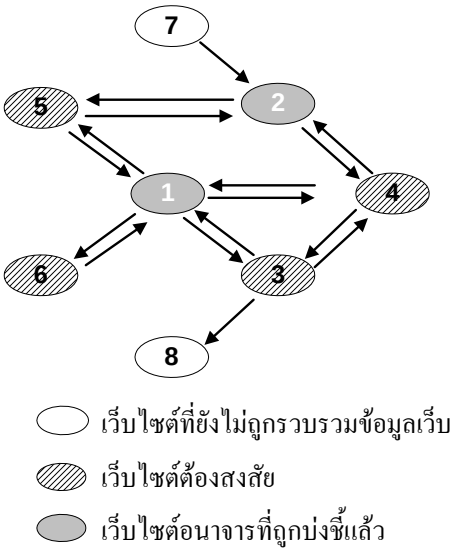
ภาพที่ 11 อัลกอริทึมฟังก์ชัน commonlink ของ LFPD

การคัดเลือก seed set ที่เหมาะสมจะทำให้การตรวจจับเว็บไซต์อนาจารเป็นไปอย่างมีประสิทธิภาพ ดังนั้นเราจึงใช้ผู้เชี่ยวชาญเป็นผู้คัดเลือก seed set โดยมีปัจจัยในการพิจารณาเว็บไซต์ที่เหมาะสมที่จะใช้เป็น seed set สองประการ ได้แก่

1. การมีเส้นเชื่อมชี้ออกไปยังเว็บไซต์อนาจารอื่น เนื่องจาก LFPD รวบรวมข้อมูลเว็บโดยติดตามเส้นเชื่อมที่ชี้ออกจาก seed set หาก seed set ไม่มีเส้นเชื่อมชี้ออกไปยังเว็บไซต์อนาจารอื่นก็จะไม่สามารถรวบรวมข้อมูลเว็บไซต์อนาจารได้ ดังนั้นจึงควรเลือก seed set จากเว็บไซต์

อนาจารที่มีเส้นเชื่อมชี้ออกไปยังเว็บไซต์อนาจารอื่นเป็นจำนวนมาก เพื่อให้ LFPD มีโอกาสรวบรวมข้อมูลเว็บไซต์อนาจารได้มากยิ่งขึ้นในจำนวนรอบการทำงานที่เท่ากัน

2. การมีเส้นเชื่อมชี้ออกไปยังกลุ่มเว็บไซต์อนาจารที่ถูกเลือกเป็น seed set ด้วยกัน เนื่องจากลักษณะของเว็บไซต์อนาจารมักมีการเชื่อมโยงกันเป็นกลุ่ม การเลือกเว็บไซต์อนาจารที่มีเส้นเชื่อมชี้ไปยัง seed set ด้วยกันจึงเป็นการดำเนินการเพื่อให้มั่นใจได้ว่าเว็บไซต์อนาจารที่ผู้เชี่ยวชาญเลือกเป็น seed set นั้นเป็นเว็บไซต์อนาจารที่อยู่ภายในกลุ่มของเว็บไซต์อนาจารดังกล่าว

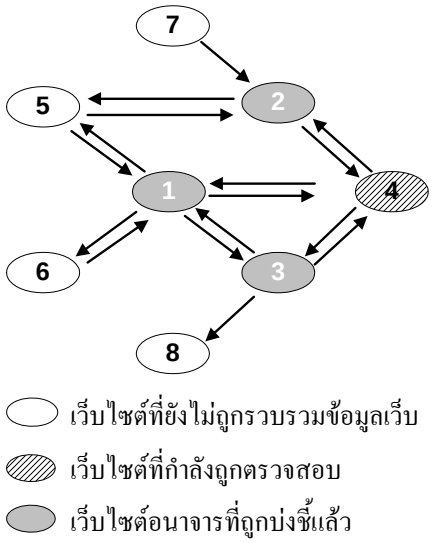


ภาพที่ 12 เว็บไซต์จำนวน 8 เว็บไซต์ และการรวบรวมเว็บไซต์ต้องสงสัย (SPS)

เว็บไซต์ต้องสงสัย (suspect pornographic site, SPS) เนื่องจากเว็บไซต์อนาจารอาจมีเส้นเชื่อมชี้ไปยังเว็บไซต์ทั่วไป หากเรารวบรวมเว็บไซต์ทั้งหมดที่พบ ก็อาจรวบรวมข้อมูลเว็บไซต์อนาจารและเว็บไซต์ทั่วไปไปพร้อมกัน ทำให้สูญเสียประสิทธิภาพในการรวบรวมข้อมูล ดังนั้นเราจึงจะรวบรวมข้อมูลจากเว็บไซต์ที่มีแนวโน้มว่าน่าจะเป็นเว็บไซต์อนาจาร โดยถือว่าเว็บไซต์ได้รับเส้นเชื่อมชี้ออกมาจากเว็บไซต์อนาจารเป็นเว็บไซต์ที่มีแนวโน้มที่จะเป็นเว็บไซต์อนาจาร หรือเรียกว่าเว็บไซต์ต้องสงสัย (suspect pornographic site, SPS) จากภาพที่ 12 ถ้าเว็บไซต์ 1 และ 2 ถูกระบุว่าเป็นเว็บไซต์อนาจาร เมื่อติดตามเส้นเชื่อมที่ชี้ออกจากเว็บไซต์ 1 และ 2 จะพบเว็บไซต์ 3,4, 5 และ 6 ซึ่งจะเป็น SPS ต่อมาเมื่อ LFPD ทำงานในรอบถัดไปและเว็บไซต์ 3 ยังไม่ถูกระบุว่าเป็นเว็บไซต์อนาจาร เมื่อติดตามเส้นเชื่อมที่ชี้ออกจากเว็บไซต์ 3 จะพบเว็บไซต์ 8 แต่กรณีนี้เว็บไซต์ 8 จะยังไม่ถูกกำหนดให้เป็น SPS จนกว่าเว็บไซต์ 3 จะถูกระบุให้เป็นเว็บไซต์อนาจาร

สำหรับงานวิจัยนี้ เว็บไซต์ที่จะถูกกำหนดให้เป็น SPS ได้แก่ เว็บไซต์ที่ได้รับเส้นเชื่อมที่ชี้มาจากเว็บไซต์อนาจารอย่างน้อย 1 เส้น

การรวบรวมข้อมูลเว็บ (Collected data) เป็นการเก็บรวบรวมเอกสารอิเล็กทรอนิกส์ประเภท HTML จาก seed set และ SPS ซึ่งเริ่มต้นโดยการอ่านเว็บเพจที่ปรากฏใน seed set และ SPS จากนั้นตรวจสอบดูว่าภายในเว็บเพจนั้นมีไฮเปอร์ลิงก์ชี้ไปยัง URL ใดบ้าง หากพบ URL ที่เป็นเอกสารอิเล็กทรอนิกล์ประเภท HTML และยังไม่เคยอ่านพบมาก่อนก็จะเก็บ URL นั้นไว้ และจะทำการอ่านเว็บเพจนั้นอีกในการทำงานรอบถัดไป นอกจากนี้เรายังได้กำหนดกฎให้รวบรวมข้อมูลเว็บในแต่ละเว็บไซต์ลึกลงไปไม่เกินสามระดับ เพื่อประหยัดเวลาและทรัพยากรในการรวบรวมข้อมูลเว็บ เนื่องจากผลการตรวจสอบเว็บไซต์อนาจารเบื้องต้นพบว่าเส้นเชื่อมที่ชี้ไปยังกลุ่มเว็บไซต์อนาจาร จะตรวจพบได้ในเว็บเพจหน้าแรกหรือหน้าถัดไป



ภาพที่ 13 เว็บไซต์จำนวน 8 เว็บไซต์ และการตรวจจับเว็บไซต์อนาจาร

การตรวจจับเว็บไซต์อนาจาร (Detected porn site) กระทำโดยการตรวจหาเว็บไซต์ที่มีเส้นเชื่อมชี้ไปและกลับ (reciprocal link หรือ common link) ร่วมกับเว็บไซต์อนาจารซึ่งถูกตรวจจับและเก็บรายชื่อไว้ใน seed set ทั้งนี้การตรวจจับโดยใช้เส้นเชื่อมประเภท common link แทนการใช้การใช้เส้นเชื่อมชี้เข้า (inlink) หรือเส้นเชื่อมชี้ออก (outlink) เพียงอย่างเดียว ก็เพื่อลดการตรวจจับผิดพลาดกรณีที่เว็บไซต์อนาจารมีเส้นเชื่อมชี้ออกไปยังเว็บไซต์ทั่วไปหรือกรณีที่เว็บไซต์ทั่วไปมีเส้นเชื่อมชี้เข้ามายังเว็บไซต์อนาจารเพียงอย่างเดียว

จากภาพที่ 13 หากเว็บไซต์ 1, 2 และ 3 เป็นเว็บไซต์อนาจารที่ถูกตรวจจับแล้ว และเว็บไซต์ 4 เป็นเว็บไซต์ที่กำลังถูกตรวจสอบ เห็นได้ว่าเว็บไซต์ 4 มี common link หรือมีเส้นเชื่อมชี้เข้าและออกร่วมกับเว็บไซต์ 1, 2 และ 3 หรือเป็น common node กับเว็บไซต์ 1, 2 และ 3 หากเรากำหนดขีดต่ำสุด(Threshold) T ในการตัดสินว่าเว็บไซต์ใดจะเป็นเว็บไซต์อนาจารเท่ากับสาม ดังนั้นเว็บไซต์ 4 จะถูกพิจารณาว่าเป็นเว็บไซต์อนาจาร สำหรับเว็บไซต์ 5 และเว็บไซต์ 6 ไม่เป็นเว็บไซต์อนาจาร เนื่องจากมี common node ร่วมกันเว็บไซต์อนาจารน้อยกว่าสาม ส่วนเว็บไซต์ 7 และ 8 ก็ไม่เป็นเว็บไซต์อนาจารเช่นกัน เนื่องจากไม่มี common link ร่วมกับเว็บไซต์อนาจาร

3. วิธีการทดลอง

เราทำการทดลองโดยผู้วิจัยได้กำหนด seed set เพื่อใช้เริ่มต้นการทดลองจำนวน 8 เว็บไซต์ จากนั้นจะให้ LFPD ทำงานไม่น้อยกว่า 3 รอบ และเปรียบเทียบผลการทำงานของ LFPD ที่ค่าขีดเริ่มเปลี่ยน (Threshold) T ต่างกัน เพื่อทดสอบความสามารถในการตรวจจับของ LFPD และทดสอบหาค่าขีดเริ่มเปลี่ยนที่เหมาะสม โดยเว็บไซต์ทั้ง 8 เว็บไซต์ ที่ผู้วิจัยได้คัดเลือกเพื่อใช้เป็น seed set เป็นเว็บไซต์อนาจารที่เหมาะสมสำหรับใช้เป็น seed set ตามสมมุติฐานที่กำหนด โดยมีคุณสมบัติได้แก่ การมีเส้นเชื่อมชี้ออกไปยังเว็บไซต์อนาจารอื่นเป็นจำนวนมาก และมีเส้นเชื่อมชี้กลับมายังกลุ่มเว็บไซต์อนาจารที่เลือกตามตารางที่ 1 และ 2

สาเหตุในการกำหนด seed set สำหรับเริ่มต้นการทำงานของ LFPD จำนวน 8 เว็บไซต์ นั้นเนื่องจากปกติแล้วการใช้ seed set เป็นจำนวนมาก และกระจายตัวในหลายๆ กลุ่มของเว็บไซต์อนาจารจะส่งผลดีต่อการทำงานของ LFPD อย่างไรก็ตามหากเราให้ผู้เชี่ยวชาญตรวจหาเว็บไซต์อนาจารเป็นจำนวนมากก็จะเป็นการสิ้นเปลืองเวลาและแรงงานของผู้เชี่ยวชาญ ซึ่งผิดไปจากวัตถุประสงค์ที่กำหนดไว้ จากการทดลองเบื้องต้นเราใช้ seed set ซึ่งเป็นเว็บไซต์ในกลุ่มเดียวกันจำนวน 5 เว็บไซต์ ปรากฏว่าสามารถตรวจจับเว็บไซต์อนาจารได้เป็นที่น่าพอใจโดยใช้ค่า T 1 ถึง 4 ดังนั้นในการทดลองครั้งนี้ เพื่อให้การตรวจจับเว็บไซต์อนาจารกว้างขวางมากยิ่งขึ้น เราจึงปรับจำนวน seed set ที่ใช้ โดยเพิ่มเว็บไซต์อีกหนึ่งกลุ่มจำนวน 3 เว็บไซต์ รวมเป็นเป็น 8 เว็บไซต์

ตารางที่ 1 รายชื่อเว็บไซต์ที่ใช้เป็น seed set

รายชื่อเว็บไซต์	จำนวนเว็บไซต์ที่ได้รับเส้นเชื่อมขี้ออก
twilightsex.com	450
Sunporno.com	606
ah-me.com	321
call-kelly.com	64
madthumbs.com	185
Doggielist.com	185
fuckingthumb.com	143
freesexportal.net	171

ตารางที่ 2 ความสัมพันธ์ระหว่างเว็บไซต์ที่ใช้เป็น seed set

รายชื่อเว็บไซต์	เว็บไซต์ที่เป็น seed set ได้รับเส้นเชื่อมขี้ออก
twilightsex.com	Sunporno.com, ah-me.com, call-kelly.com, madthumbs.com
Sunporno.com	twilightsex.com, ah-me.com, call-kelly.com, madthumbs.com
ah-me.com	twilightsex.com, sunporno.com
call-kelly.com	twilightsex.com, sunporno.com, ah-me.com, madthumbs.com
madthumbs.com	twilightsex.com, sunporno.com, call-kelly.com
Doggielist.com	fuckingthumb.com
fuckingthumb.com	Doggielist.com, freesexportal.net
freesexportal.net	fuckingthumb.com

4. การประเมินผลการทดลอง

การประเมินผลการทดลองกระทำโดยการเปรียบเทียบความแม่นยำ (precision) และอัตราการตรวจจับ (Detection Rate) ของ LFPD ที่ T ต่างๆ โดยข้อมูลที่ใช้ตรวจสอบผลการทำงานของ LFPD นั้น นำมาจากรายชื่อเว็บไซต์อนาจารที่เผยแพร่ในเว็บไซต์ URLBlasklist.com (URLBlacklist.com, 2008) ซึ่งเป็นเว็บไซต์ที่ให้บริการเกี่ยวกับรายชื่อเว็บไซต์ในกลุ่มต่างๆ รวมทั้ง

กลุ่มเว็บไซต์อนาจาร โดยเป็นข้อมูลที่ดาวน์โหลดเมื่อเมื่อวันที่ 14 กันยายน 2551 มีรายชื่อเว็บไซต์อนาจารทั้งสิ้นจำนวน 877,187 เว็บไซต์ ทั้งนี้เว็บไซต์ URLBlacklist.com ระบุว่าที่มาของข้อมูลรายชื่อเว็บไซต์ที่ให้บริการนั้นเป็นการรายงานจากผู้ใช้และตรวจสอบยืนยันโดยผู้เชี่ยวชาญอีกครั้งหนึ่ง (human verified user submissions)

การประเมินผลการทดลองกระทำโดยใช้สูตรดังนี้

$$Precision = \frac{ACC}{AC} \quad (1)$$

$$Detection\ Rate = \frac{ACC}{TC} \quad (2)$$

โดยกำหนดให้

Algorithm Confirmed Correctly (A*CC*) = เว็บไซต์อนาจารที่อัลกอริทึมตรวจจับได้ถูกต้อง (ตรงกับ Blacklist)

Algorithm Confirmed (A*C*) = เว็บไซต์อนาจารทั้งหมดที่ตรวจจับได้โดยอัลกอริทึม

Total Collected (*TC*) = เว็บไซต์ที่รวบรวมได้ทั้งหมด (seed set and SPS)

ค่าความแม่นยำ (Precision) บ่งบอกถึงความถูกต้องของข้อมูลเว็บไซต์อนาจารที่ถูกตรวจจับจาก LFPD และอัตราการตรวจจับ (Detection Rate) บ่งบอกถึงความสามารถในการรวบรวมข้อมูลเว็บโดยเฉพาะในส่วนที่เป็นเว็บไซต์อนาจาร โดยเปรียบเทียบระหว่างจำนวนเว็บไซต์อนาจารที่ตรวจจับได้ถูกต้องกับจำนวนเว็บไซต์ที่รวบรวมข้อมูลมาทั้งหมด หากจำนวนเว็บไซต์ที่รวบรวมข้อมูลมาทั้งหมดใกล้เคียงกับจำนวนเว็บไซต์อนาจารที่ตรวจจับได้ แสดงว่า LFPD มีประสิทธิภาพเนื่องจากสามารถตรวจจับเว็บไซต์อนาจารได้โดยไม่จำเป็นต้องรวบรวมข้อมูลเว็บที่ไม่เกี่ยวข้องเป็นจำนวนมาก

ผลและวิจารณ์

เนื้อหาในส่วนนี้จะกล่าวถึงผลการทำงานของ LFPD ที่ทำงานตามปกติ และเปรียบเทียบกับการทำงานภายใต้เงื่อนไขอื่น พร้อมทั้งวิจารณ์ผลการทดลองที่ได้รับว่ามีความสอดคล้องหรือแตกต่างจากผลที่คาดไว้หรือไม่อย่างไร

เราทำการทดลองโดยรวบรวมข้อมูลเว็บในช่วงเดือนกันยายน – ตุลาคม 2551 และทำการทดลองสี่ประการ ได้แก่ การทำงานของ LFPD ตามปกติ, การทำงานของ LFPD โดยไม่ใช้ SPS, การทำงานของ LFPD ที่ตรวจจับด้วย colink และการทำงานของ LFPD ที่ใช้ seed set ไม่เหมาะสม สรุปผลการทดลองได้ดังนี้

1. ผลการทำงานของ LFPD

ผลการทดลองประการแรกได้แก่ ผลการทำงานของ LFPD ตามปกติ ซึ่งตรวจจับเว็บไซต์อนาจารโดยการนับจำนวน common node ที่เกิดจาก common link ซึ่งมีค่าเท่ากับหรือสูงกว่าขีดเริ่มเปลี่ยน (Threshold) T ที่กำหนด โดยค่าของ T ที่สามารถใช้ได้มีค่าตั้งแต่ 1 ถึง 4

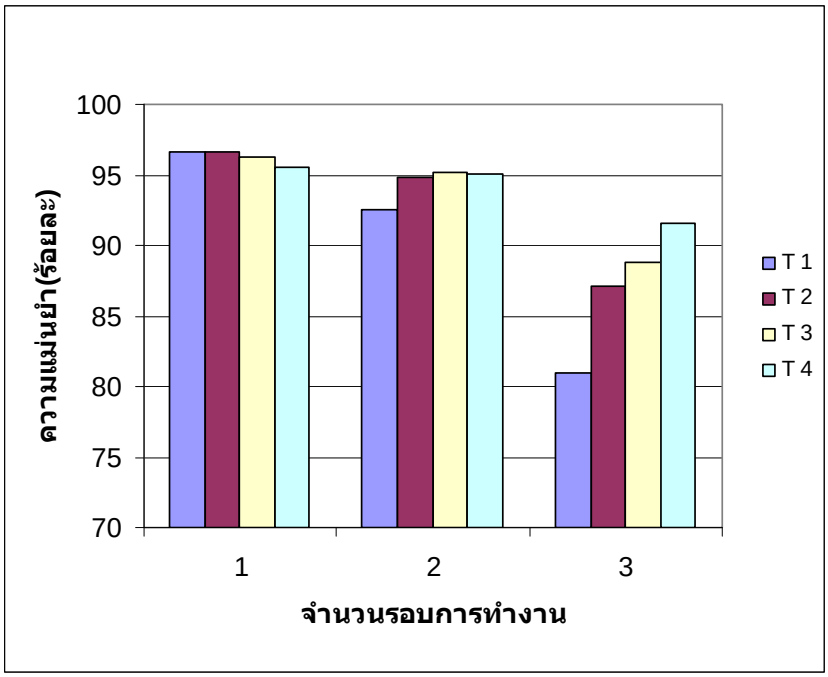
ตารางที่ 3 ผลการทำงานของ LFPD

รายการ	ผลการทำงาน		
	รอบที่ 1	**รอบที่ 2**	**รอบที่ 3**
T 1			
เว็บไซต์ที่ตรวจจับได้ทั้งหมด	445	10,264	54,518
เว็บไซต์ที่ตรวจจับได้และตรงกับ Blacklist	430	9,500	44,123
Seed set และ SPS ที่ตรงกับ Blacklist	1,229	19,259	94,329
Seed set และ SPS	1,360	22,981	143,143
ความแม่นยำ(Precision)(ร้อยละ)	96.63	92.56	80.93
อัตราการตรวจจับ (Detection Rate) (ร้อยละ)	31.62	41.34	30.82

ตารางที่ 3 (ต่อ)

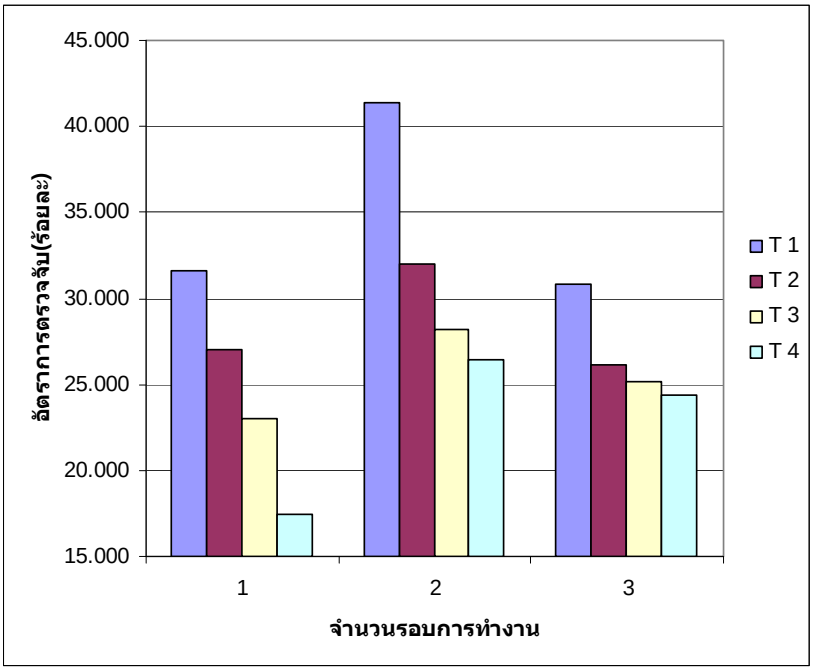
รายการ	ผลการทำงาน		
	รอบที่ 1	**รอบที่ 2**	**รอบที่ 3**
T 2			
เว็บไซต์ที่ตรวจจับได้ทั้งหมด	381	7,401	37,069
เว็บไซต์ที่ตรวจจับได้และตรงกับ Blacklist	368	7,022	32,298
Seed set และ SPS ที่ตรงกับ Blacklist	1,229	18,422	84,786
Seed set และ SPS	1,360	21,919	123,359
ความแม่นยำ(Precision)(ร้อยละ)	96.59	94.88	87.13
อัตราการตรวจจับ (Detection Rate) (ร้อยละ)	27.06	32.04	26.18
T 3			
เว็บไซต์ที่ตรวจจับได้ทั้งหมด	325	5,820	29,916
เว็บไซต์ที่ตรวจจับได้และตรงกับ Blacklist	313	5,539	26,549
Seed set และ SPS ที่ตรงกับ Blacklist	1,229	16,433	77,794
Seed set และ SPS	1,360	19,660	105,709
ความแม่นยำ(Precision)(ร้อยละ)	96.31	95.17	88.75
อัตราการตรวจจับ (Detection Rate) (ร้อยละ)	23.01	28.17	25.12
T 4			
เว็บไซต์ที่ตรวจจับได้ทั้งหมด	248	3,564	20,440
เว็บไซต์ที่ตรวจจับได้และตรงกับ Blacklist	237	3,388	18,725
Seed set และ SPS ที่ตรงกับ Blacklist	1,229	11,110	58,746
Seed set และ SPS	1,360	12,840	76,651
ความแม่นยำ(Precision)(ร้อยละ)	95.56	95.06	91.61
อัตราการตรวจจับ (Detection Rate) (ร้อยละ)	17.43	26.39	24.43

จากผลการทดลองที่ปรากฏในตารางที่ 3 พบว่าเมื่อให้ LFPD ทำงานไปครบ 3 รอบ เราสามารถตรวจจับเว็บไซต์อนาจารได้ตั้งแต่ 18,725 ถึง 44,123 เว็บไซต์ ขึ้นอยู่กับ T ที่ใช้ โดยความแม่นยำจะมีค่าเพิ่มขึ้นเมื่อ T มีค่าสูงขึ้น ส่วนอัตราการตรวจจับจะมีค่าเพิ่มขึ้นเมื่อ T มีค่าลดลง สำหรับจำนวนเว็บไซต์ที่ตรวจจับได้ก็จะมีค่าเพิ่มขึ้นเมื่อ T มีค่าลดลงด้วยเช่นเดียวกัน



ภาพที่ 14 เปรียบเทียบความแม่นยำของ LFPD

เมื่อนำค่าความแม่นยำมาแสดงด้วยกราฟตามภาพที่ 14 เห็นได้ว่าผลการทำงานในรอบที่ 1 ค่าความแม่นยำจากการใช้ T 1 ถึง 4 มีค่าสูงกว่าร้อยละ 95 โดยเมื่อใช้ T 1 และ 2 จะมีค่าความแม่นยำสูงกว่า T 3 และ 4 อยู่เล็กน้อย แต่เมื่อทำงานต่อไปในรอบที่ 2 ค่าความแม่นยำของการใช้ T 1 และ 2 จะเริ่มลดลงจนมีค่าต่ำกว่าการใช้ T 3 และ 4 กระทั่งในรอบที่ 3 พบว่าค่าความแม่นยำจะลดลงตามค่าของ T ที่ใช้ตามลำดับ จนเกือบทั้งหมดมีความแม่นยำต่ำกว่าร้อยละ 90



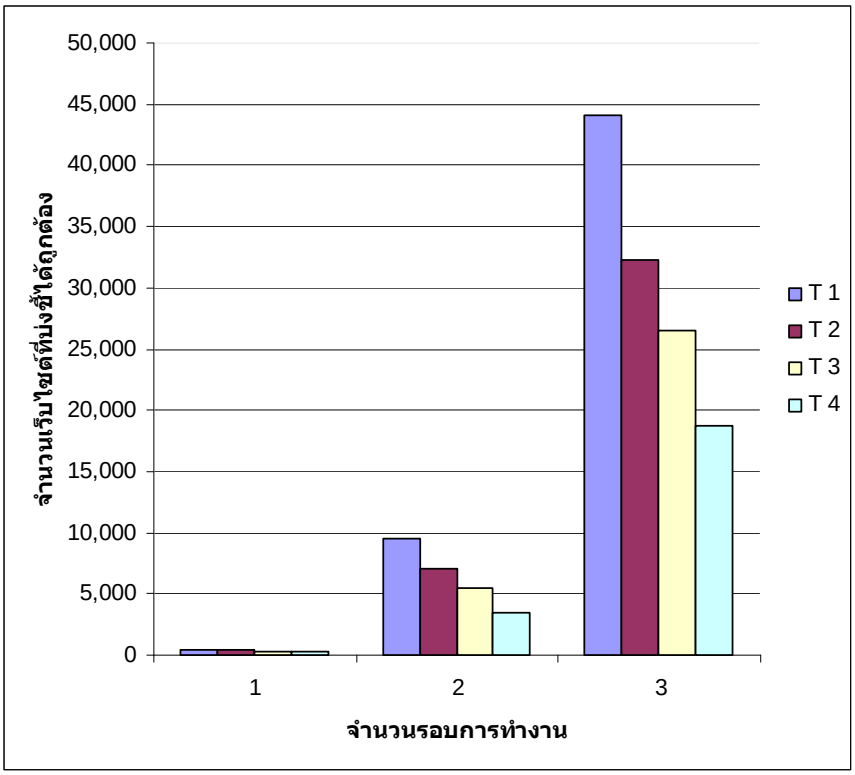
ภาพที่ 15 เปรียบเทียบอัตราการตรวจจับของ LFPD

จากภาพที่ 15 เห็นได้ว่าอัตราการตรวจจับการทำงานของ LFPD จะลดต่ำลงตามค่า T ที่เพิ่มขึ้น โดยมีอัตราการตรวจจับสูงสุดเมื่อ T เท่ากับ 1 และมีค่าต่ำที่สุดเมื่อ T เท่ากับ 4 นอกจากนี้อัตราการตรวจจับยังเปลี่ยนแปลงตามรอบการทำงาน โดยในรอบที่ 2 เป็นจุดที่ LFPD ให้อัตราการตรวจจับสูงที่สุด

เมื่อวิเคราะห์ผลการทดลองข้างต้นพบว่า ปัญหาหลักที่ทำให้ LFPD มีค่าความแม่นยำลดลงตามวงรอบการทำงานที่เพิ่มขึ้นเนื่องมาจากในการทำงานของ LFPD นั้น เว็บไซต์อนาจารที่ตรวจจับได้ในรอบก่อนหน้าจะถูกนำกลับมารวมกับ seed set ที่มีอยู่เดิม เพื่อใช้เป็น seed set ใหม่สำหรับใช้ทำงานในรอบถัดไป ดังนั้นความแม่นยำในแต่ละรอบการทำงานของ LFPD จึงขึ้นอยู่กับผลการตรวจจับในรอบก่อนหน้าด้วย ซึ่งในแต่ละรอบการทำงานของ LFPD เราพบว่ามีเว็บไซต์ที่ตรวจจับผิดพลาดอยู่จำนวนหนึ่ง โดยเป็นการระบุเว็บไซต์ทั่วไปให้เป็นเว็บไซต์อนาจาร การตรวจจับผิดพลาดดังกล่าวทำให้มีเว็บไซต์ทั่วไปสะสมใน seed set เพิ่มมากขึ้นในแต่ละรอบการทำงาน ซึ่งจะทำให้ความแม่นยำของ LFPD ลดลง ทั้งนี้สาเหตุการตรวจจับผิดพลาดข้างต้นเกิดจากการที่เว็บไซต์ทั่วไปดังกล่าวมีเส้นเชื่อมชี้ไปและกลับร่วมกับเว็บไซต์อนาจาร ซึ่งไม่สอดคล้องกับสมมุติฐานที่คาดการณ์ไว้ เมื่อตรวจสอบเว็บไซต์ทั่วไปที่ถูกตรวจจับโดยมีเส้นเชื่อมชี้ไปและกลับร่วมกับ

เว็บไซต์อนาจารเราพบว่าเป็นเส้นเชื่อมที่สร้างขึ้นโดยเจตนาเนื่องจากปรากฏอยู่ทั้งในโฮมเพจและที่เว็บเพจอื่น ทั้งนี้เว็บไซต์ทั่วไปดังกล่าวสามารถแยกเป็นกลุ่มต่างๆ เช่น กลุ่มดาวน์โหลดภาพยนตร์, กลุ่มการบริการเงินสด , กลุ่มการพนัน, กลุ่มหาคู่, กลุ่มเกมส์ เป็นต้น

นอกจากนี้ผลกระทบจากการตรวจจับผิดพลาดยังสะท้อนให้เห็นด้วยอัตราการตรวจจับที่มีค่าสูงสุดในรอบที่ 2 และลดต่ำลงในรอบที่ 3 เนื่องจากการทำงานในรอบที่ 1 ยังคงมีความแม่นยำสูง ทำให้ seed set ที่ได้มีเว็บไซต์ทั่วไปปะปนอยู่น้อย ดังนั้นการทำงานในรอบที่ 2 จึงยังคงมีอัตราการตรวจจับที่สูง แต่เมื่อทำงานต่อไปในรอบที่ 2 จะมีการสะสมเว็บไซต์ทั่วไปที่ถูกตรวจจับที่ผิดพลาดไว้ใน seed set เพิ่มมากขึ้น ทำให้ในรอบที่ 3 LFPD จึงต้องรวบรวมข้อมูลเว็บจาก seed set ที่มีเว็บไซต์ทั่วไปที่ถูกตรวจจับผิดพลาดรวมอยู่ด้วยมากขึ้น จึงส่งผลให้อัตราการตรวจจับมีค่าลดลง



ภาพที่ 16 เปรียบเทียบจำนวนเว็บไซต์ที่ LFPD ตรวจจับได้ถูกต้อง

เมื่อนำจำนวนเว็บไซต์ที่ LFPD ตรวจจับได้ถูกต้องมาแสดงด้วยกราฟตามภาพที่ 16 เห็นได้ว่าจำนวนเว็บไซต์ที่ตรวจจับได้ถูกต้องจะลดลงเมื่อ T มีค่าเพิ่มขึ้น และจำนวนเว็บไซต์ที่ตรวจจับได้ถูกต้องจะเพิ่มขึ้นตามรอบการทำงานที่เพิ่มขึ้น โดยเมื่อเปรียบเทียบผลการทำงานในรอบที่ 1 กับ

รอบที่ 2 พบว่าจำนวนเว็บไซต์ที่ตรวจจับได้ถูกต้องจะเพิ่มขึ้นเป็นอัตราส่วน 22.09, 19.08, 17.70 และ 14.30 เท่า ตามลำดับค่า T ที่ใช้ และเมื่อเปรียบเทียบผลการทำงานในรอบที่ 2 กับรอบที่ 3 จำนวนเว็บไซต์ที่ตรวจจับได้ถูกต้องจะเพิ่มขึ้นเป็นอัตราส่วน 4.64, 4.60, 4.79 และ 5.53 เท่า ตามลำดับค่า T ที่ใช้

การที่อัตราส่วนจำนวนเว็บไซต์ที่ตรวจจับได้ถูกต้องในรอบที่ 1 กับ 2 มีค่าสูงเกือบ 20 เท่า และลดลงเหลือราว 5 เท่า ในรอบที่ 2 กับ 3 ทั้งๆ ที่ยังคงมีเว็บไซต์อนาจารที่ยังตรวจไม่พบอีกเป็นจำนวนมากนั้น สาเหตุสำคัญมีสองประการ ประการแรกเป็นผลมาจากปัญหาที่กล่าวไปแล้ว ได้แก่ การตรวจจับที่ผิดพลาดทำให้มีเว็บไซต์ทั่วไปปะปนอยู่ใน seed set และ seed set มีคุณภาพต่ำลง สาเหตุประการที่สองนั้นสืบเนื่องมาจากข้อจำกัดของกลุ่ม seed set ที่ถูกเลือกเมื่อเริ่มต้นทำการทดลองซึ่งจำกัดจำนวนเพียง 8 เว็บไซต์ จึงไม่สามารถครอบคลุมเว็บไซต์อนาจารได้ทั้งหมดทุกกลุ่ม ทั้งนี้ปัญหาดังกล่าวมีลักษณะเช่นเดียวกันกับจุดอ่อนของ TrustRank ที่เป็นการยากที่จะเลือกกลุ่มของเว็บเพจดีให้เพียงพอที่จะเป็นตัวแทนของเว็บไซต์แต่ละกลุ่ม (Wu et al., 2006) นั่นเอง

ตารางที่ 4 เว็บไซต์ที่ LFPD ตรวจจับได้เปรียบเทียบกับเว็บไซต์ที่ปรากฏใน Blacklist

รายการ	**จำนวนเว็บไซต์อนาจาร**			
	T 1	**T 2**	**T 3**	**T 4**
รอบที่ 1				
เว็บไซต์ที่ตรวจจับได้ทั้งหมด	445	381	325	248
เว็บไซต์ที่ตรวจจับได้และตรงกับ Blacklist	430	368	313	237
เว็บไซต์ที่ตรวจจับได้เพิ่มเติมจาก Blacklist	15	13	12	11
เว็บไซต์ที่ตรวจจับได้เพิ่มเติมจาก Blacklist และได้รับการยืนยันจากผู้เชี่ยวชาญ(ร้อยละ)	14(93%)	12(92%)	11(91%)	10(91%)
รอบที่ 2				
เว็บไซต์ที่ตรวจจับได้ทั้งหมด	10,264	7,401	5,820	3,564
เว็บไซต์ที่ตรวจจับได้และตรงกับ Blacklist	9,500	7,022	5,539	3,388
เว็บไซต์ที่ตรวจจับได้เพิ่มเติมจาก Blacklist	764	379	281	176
เว็บไซต์ที่ตรวจจับได้เพิ่มเติมจาก Blacklist และได้รับการยืนยันจากผู้เชี่ยวชาญ(ร้อยละ)	659(86%)	347(92%)	255(91%)	161(91%)

จากข้อมูลในตารางที่ 4 พบว่ามีเว็บไซต์จำนวนหนึ่งที่ LFPD ระบุว่าเป็นเว็บไซต์อนาจาร แต่ไม่ปรากฏรายชื่อในข้อมูลเว็บไซต์อนาจารหรือ Blacklist ที่ได้จากเว็บไซต์ URLBlasklist.com เมื่อเราตรวจสอบเว็บไซต์ดังกล่าวที่ LFPD ตรวจจับได้ในการทำงานรอบที่ 1 และ 2 ซึ่งมีจำนวนทั้งสิ้น 764 เว็บไซต์ ด้วยผู้เชี่ยวชาญ พบว่าประมาณร้อยละ 90 ของเว็บไซต์อนาจารที่ตรวจจับได้โดยไม่ปรากฏรายชื่อในข้อมูลเว็บไซต์อนาจารหรือ Blacklist เป็นเว็บไซต์อนาจาร

การที่ตรวจพบเว็บไซต์อนาจารนอกเหนือจากรายชื่อที่ปรากฏใน Blacklist นั้น เนื่องมาจากเว็บไซต์อนาจารจะมีการเปลี่ยนแปลงและเพิ่มจำนวนอย่างต่อเนื่อง การจัดทำบัญชีรายชื่อเว็บไซต์โดยใช้ข้อมูลรายงานจากผู้ใช้และตรวจสอบยืนยันโดยผู้เชี่ยวชาญ (human verified user submissions) ไม่อาจติดตามการเปลี่ยนแปลงดังกล่าวได้อย่างทันท่วงที ข้อมูลที่ได้จึงไม่ทันสมัยหรือไม่ครอบคลุมรายชื่อเว็บไซต์อนาจารอย่างทั่วถึง ซึ่งจะเห็นว่าการทำงานของ LFPD ที่รวบรวมข้อมูลเว็บไซต์อนาจารโดยติดตามเส้นเชื่อมระหว่างเว็บไซต์ทำงานโดยมีประสิทธิภาพที่ดีกว่าสามารถตรวจจับเว็บไซต์อนาจารที่มีเส้นเชื่อมชี้ถึงกันได้ใกล้เคียงเวลาจริงที่สร้างหรือปรับปรุงเปลี่ยนแปลงเว็บไซต์อนาจารในช่วงที่ LFPD ทำงาน เห็นได้จากเว็บไซต์อนาจารที่ LFPD ตรวจจับได้เพิ่มเติมในการทำงานรอบที่ 2 ซึ่งมีจำนวนตั้งแต่ 161 ถึง 659 เว็บไซต์ ขึ้นอยู่กับ T ที่ใช้ สำหรับเว็บไซต์ทั่วไปที่ LFPD ระบุผิดพลาดประมาณร้อยละ 10 ของเว็บไซต์ที่ตรวจจับได้นอกเหนือจาก Blacklist เป็นเว็บไซต์ทั่วไปที่มีเส้นเชื่อมชี้ไปและกลับร่วมกับเว็บไซต์อนาจาร ซึ่งก่อให้เกิดการตรวจจับผิดพลาดตามที่กล่าวไปแล้วข้างต้น

อย่างไรก็ตาม LFPD ก็มีข้อจำกัดที่ไม่สามารถตรวจจับเว็บไซต์อนาจารที่มีลักษณะโดดเดี่ยวโดยไม่มีเส้นเชื่อมชี้เข้าหรือออกกร่วมกับเว็บไซต์อื่น (isolated) หรือเป็นเว็บไซต์อนาจารที่ไม่มีเส้นเชื่อมชี้ออก (outlink) มายังเว็บไซต์อนาจารที่เราตรวจพบ เนื่องจาก LFPD จะตรวจจับเว็บไซต์อนาจารจากเว็บไซต์ที่มีเส้นเชื่อมชี้ไปและกลับ (reciprocal link หรือ common link) รวมกับเว็บไซต์อนาจารที่ตรวจพบก่อนหน้า

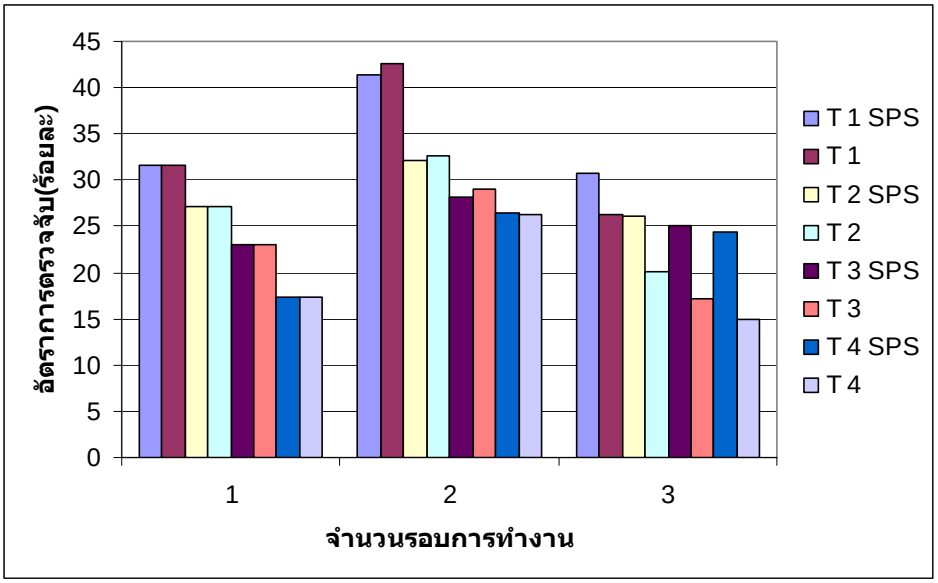
2. ผลการทำงานของ LFPD โดยไม่ใช้ SPS

ผลการทดลองประการที่สองได้แก่ ผลการทำงานของ LFPD โดยไม่ใช้ SPS แต่ยังคงตรวจจับโดยการนับจำนวน common node ที่เกิดจาก common link ซึ่งมีค่าเท่ากับหรือสูงกว่าขีดเริ่มเปลี่ยน (Threshold) T ที่กำหนด โดยขีดเริ่มเปลี่ยนที่ใช้ในการทดลองมีค่าตั้งแต่ 1 ถึง 4

เช่นเดียวกันกับกรณีแรก ทั้งนี้การเปรียบเทียบผลการทำงานของ LFPD ที่ได้รับจากการใช้และไม่ใช้ SPS จะทำให้ทราบถึงประโยชน์ในการใช้ SPS ได้ต่อไป

ตารางที่ 5 เปรียบเทียบอัตราการตรวจจับของ LFPD กรณีที่ใช้และไม่ใช้ SPS

รายการ	ใช้ SPS		ไม่ใช้ SPS	
	จำนวนเว็บไซต์ที่ตรวจจับได้ถูกต้อง	อัตราการตรวจจับ (Detection Rate) (ร้อยละ)	จำนวนเว็บไซต์ที่ตรวจจับได้ถูกต้อง	อัตราการตรวจจับ (Detection Rate) (ร้อยละ)
T 1				
รอบที่ 1	430	31.62	430	31.62
รอบที่ 2	9,500	41.34	11,097	42.58
รอบที่ 3	44,123	30.82	50,036	26.25
T 2				
รอบที่ 1	368	27.06	368	27.06
รอบที่ 2	7,022	32.04	8,497	32.60
รอบที่ 3	32,298	26.18	38,156	20.01
T 3				
รอบที่ 1	313	23.01	313	23.01
รอบที่ 2	5,539	28.17	7,545	28.95
รอบที่ 3	26,549	25.12	32,731	17.17
T 4				
รอบที่ 1	237	17.43	237	17.43
รอบที่ 2	3,388	26.39	6,840	26.24
รอบที่ 3	18,725	24.43	28,327	14.86



ภาพที่ 17 เปรียบเทียบอัตราการตรวจจับของ LFPD ระหว่างการใช้และไม่ใช้ SPS

เมื่อพิจารณาอัตราการตรวจจับของ LFPD เมื่อใช้และไม่ใช้ SPS ตามข้อมูลที่ปรากฏในตารางที่ 5 และกราฟตามภาพที่ 17 เห็นได้ว่าการทำงานของ LFPD ที่ใช้และมิได้ใช้ SPS จะมีค่าใกล้เคียงกันในรอบที่ 2 แต่เมื่อทำงานต่อไปในรอบที่ 3 อัตราการตรวจจับของ LFPD ที่ใช้ SPS จะมีค่าสูงกว่าเมื่อไม่ใช้ SPS อย่างเห็นได้ชัด

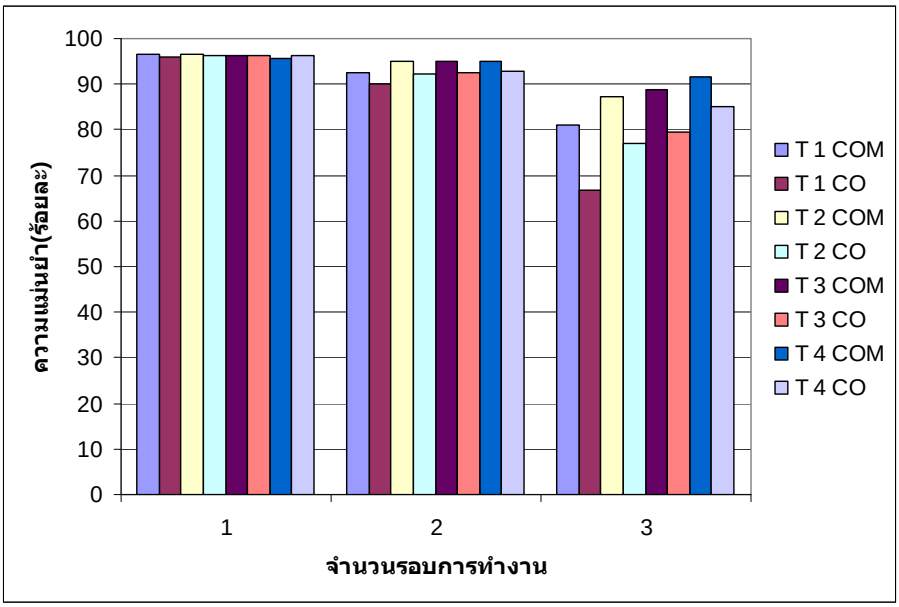
การที่ผลการทำงานของ LFPD ที่ใช้และไม่ใช้ SPS ให้อัตราการตรวจจับใกล้เคียงกันในการทำงานรอบที่ 2 เนื่องจากเว็บไซต์ทั้งหมดที่ตรวจพบในรอบที่ 1 เป็นเว็บไซต์อนาจารที่มีเส้นเชื่อมชี้มากจาก seed set ที่ผู้เชี่ยวชาญกำหนดสำหรับเริ่มต้นการทำงานโดยตรง จึงมีเส้นเชื่อมชี้ไปยังเว็บไซต์อนาจารด้วยกันเป็นส่วนใหญ่ เมื่อไม่ใช้ SPS จึงเป็นการรวบรวมข้อมูลเว็บจากเว็บไซต์ที่ได้รับเส้นเชื่อมที่ชี้จากเว็บไซต์ที่เป็น seed set ในรอบที่ 1 ทั้งหมด ทำให้สามารถตรวจพบเว็บไซต์อนาจารได้เป็นจำนวนมาก ซึ่งหากใช้ SPS ก็จะเป็นการรวบรวมข้อมูลเว็บเฉพาะเว็บไซต์ที่ถูกตรวจจับแล้วเท่านั้น อย่างไรก็ตามเมื่อทำงานต่อไปในรอบที่ 3 ซึ่งเริ่มที่จะมีเว็บไซต์ทั่วไปที่ได้รับเส้นเชื่อมชี้ออกจากเว็บไซต์อนาจารเพิ่มมากขึ้น จึงปรากฏว่า LFPD ที่มิได้ใช้ SPS เริ่มมีอัตราการตรวจจับต่ำกว่าการใช้ SPS เนื่องจากว่าเมื่อไม่ใช้ SPS แล้ว LFPD ก็จะทำการรวบรวมข้อมูลเว็บจากเว็บไซต์ทั่วไปที่มิได้ถูกตรวจจับไปอย่างต่อเนื่อง ในขณะที่ LFPD ที่ใช้ SPS จะยังคงรวบรวมข้อมูลเว็บจากเว็บไซต์อนาจารที่ถูกตรวจจับแล้วเท่านั้น

3. ผลการทำงานของ LFPD ที่ตรวจจับด้วย colink

ผลการทดลองประการที่สามได้แก่ ผลการทำงานของ LFPD ที่ทำงานตามปกติเปรียบเทียบกับ LFPD ที่ปรับใช้การตรวจจับด้วย colink ซึ่งจะระบุว่าเว็บไซต์ใดเป็นเว็บไซต์อนาจาร ก็ต่อเมื่อเว็บไซต์นั้นมีเส้นเชื่อมชี้กลับมายังเว็บไซต์อนาจารที่ตรวจจับได้แล้วเป็นจำนวนเท่ากับหรือมากกว่าค่า T ที่กำหนด การเปรียบเทียบผลการทำงานของ LFPD ที่ทำงานตามปกติกับ LFPD ที่บ่งชี้ด้วย colink จะทำให้ทราบถึงความเป็นไปได้ในการใช้บ่งชี้ด้วยโครงสร้างเส้นเชื่อมในลักษณะอื่นต่อไป

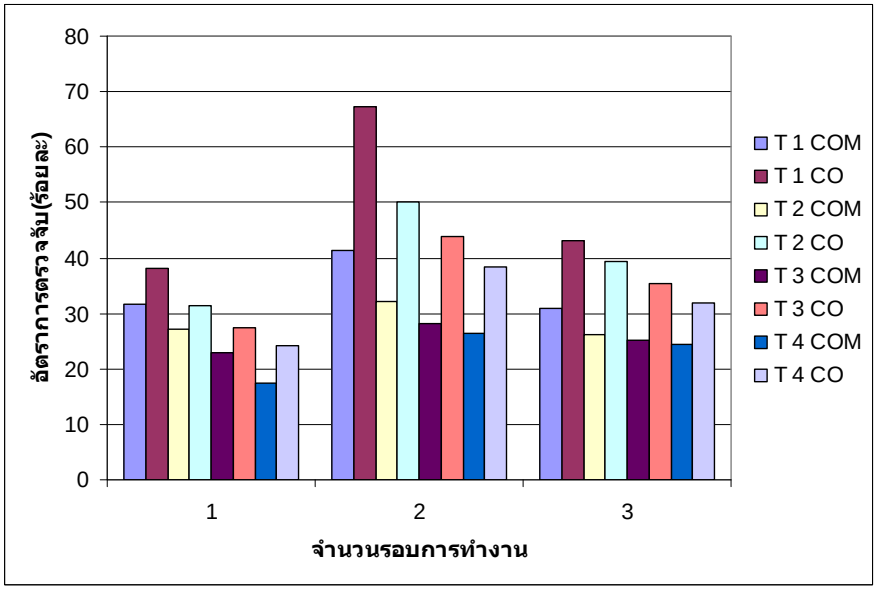
ตารางที่ 6 เปรียบเทียบความแม่นยำในการทำงานของ LFPD ระหว่าง common link กับ colink

รายการ	common link		Colink	
	จำนวนเว็บไซต์ที่ตรวจจับได้ถูกต้อง	ความแม่นยำ (precision) (ร้อยละ)	จำนวนเว็บไซต์ที่ตรวจจับได้ถูกต้อง	ความแม่นยำ (precision) (ร้อยละ)
T 1				
รอบที่ 1	430	96.63	519	96.11
รอบที่ 2	9,500	92.56	16,341	90.19
รอบที่ 3	44,123	80.93	77,576	66.87
T 2				
รอบที่ 1	368	96.59	428	96.40
รอบที่ 2	7,022	94.88	11,605	92.29
รอบที่ 3	32,298	87.13	59,568	76.92
T 3				
รอบที่ 1	313	96.31	374	96.39
รอบที่ 2	5,539	95.17	9,899	92.44
รอบที่ 3	26,549	88.75	50,813	79.38
T 4				
รอบที่ 1	237	95.56	328	96.19
รอบที่ 2	3,388	95.06	7,868	92.71
รอบที่ 3	18,725	91.61	40,899	84.97



ภาพที่ 18 เปรียบเทียบความแม่นยำของ LFPD ระหว่าง common link กับ colink

เมื่อนำค่าความแม่นยำในการทำงานของ LFPD ที่ทำงานตามปกติเปรียบเทียบกับ LFPD ที่ตรวจจับด้วย colink มาแสดงด้วยกราฟตามภาพที่ 18 เห็นได้ว่าความแม่นยำในการทำงานของ LFPD ที่ตรวจจับด้วย common link ตามปกติจะมีค่าสูงกว่า LFPD ที่ตรวจจับด้วย colink



ภาพที่ 19 เปรียบเทียบอัตราการตรวจจับของ LFPD ระหว่าง common link กับ colink

แต่เมื่อนำค่าอัตราการตรวจจับของ LFPD ที่ทำงานตามปกติกับ LFPD ที่ตรวจจับด้วย colink มาแสดงด้วยกราฟตามภาพที่ 19 เห็นได้ว่าอัตราการตรวจจับของ LFPD ที่ตรวจจับด้วย common link ตามปกติจะมีค่าต่ำกว่า LFPD ที่ตรวจจับด้วย colink

ผลการเปรียบเทียบการทำงานของ LFPD ที่ตรวจจับด้วย common link และ colink แสดงให้เห็นว่า colink ก็สามารถใช้ตรวจจับเว็บไซต์อนาจารได้เช่นกัน ทั้งนี้การที่ตรวจจับด้วย common link ให้ความแม่นยำสูงกว่า colink ในช่วงระหว่างร้อยละ 0.52 ถึง 6.64 ขึ้นอยู่กับจำนวนรอบการทำงานและค่า T ที่ใช้ เนื่องจากการที่เว็บไซต์มีเส้นเชื่อมชี้ไปและกลับระหว่างกันย่อมหมายถึงการให้ความสำคัญระหว่างกันสูงกว่าการมีเส้นเชื่อมชี้ไปหรือกลับเพียงอย่างเดียว แต่การได้มาซึ่งความแม่นยำที่สูงกว่านั้นก็ต้องแลกด้วยอัตราการตรวจจับที่ลดต่ำลงไป อย่างไรก็ตามเมื่อพิจารณาถึงจำนวนเว็บไซต์อนาจารที่ตรวจจับได้ถูกต้องแล้วพบว่าที่จำนวนเว็บไซต์ที่ตรวจจับได้ถูกต้องใกล้เคียงกันนั้นทั้ง common link และ colink ก็ให้ค่าความแม่นยำที่ใกล้เคียงกัน ดังนั้นเราจึงยังไม่สามารถชี้ชัดลงไปได้ว่าการตรวจจับแบบใดให้ผลลัพธ์ที่ดีกว่ากัน

4. ผลการทำงานของ LFPD ที่ใช้ seed set ไม่เหมาะสม

ผลการทดลองประการที่สี่เป็นการเปรียบเทียบผลการทำงานของ LFPD โดยใช้ seed set ที่เหมาะสมตามตารางที่ 1 และ 2 เปรียบเทียบกับการใช้ seed set ที่ไม่เหมาะสมตามตารางที่ 7 และ 8

ตารางที่ 7 รายชื่อเว็บไซต์ที่ใช้เป็น seed set ที่ไม่เหมาะสม

รายชื่อเว็บไซต์	จำนวนเว็บไซต์ที่ได้รับเส้นเชื่อมชี้ออก
0-dix.com	26
1sluty.com	55
Arabporn.info	54
bigboobsdiary.com	118
keezmovies.com	10
Pornhub.com	17
milf-addict.com	69
yourpornclips.com	46

ตารางที่ 8 ความสัมพันธ์ระหว่างเว็บไซต์ที่ใช้เป็น seed set ที่ไม่เหมาะสม

รายชื่อเว็บไซต์	**เว็บไซต์ที่เป็น seed set ได้รับเส้นเชื่อมขี้ออก**
0-dix.com	1sluty.com, arabporn.info, bigboobsdiary.com, keezmovies.com, pornhub.com, milf-addict.com, yourpornclips.com
1sluty.com	0-dix.com
Arabporn.info	0-dix.com
bigboobsdiary.com	0-dix.com
keezmovies.com	
Pornhub.com	0-dix.com
Milf-addict.com	0-dix.com
yourpornclips.com	0-dix.com

ตารางที่ 9 เปรียบเทียบผลการทำงานในรอบที่ 1 ของ LFPD เมื่อกำหนด seed set ที่เหมาะสมและไม่เหมาะสม

รายการ	**จำนวนเว็บไซต์ที่ตรวจจับได้ถูกต้องในการทำงานรอบที่ 1**	
	ใช้ seed set ที่เหมาะสม	**ใช้ seed set ที่ไม่เหมาะสม**
T 1	430	89
T 2	368	50
T 3	313	0
T 4	237	0

จากผลการทดลองซึ่งปรากฏตามตารางที่ 9 เห็นได้ว่าการคัดเลือก seed set ที่เหมาะสม โดยมีคุณสมบัติได้แก่ การเป็นเว็บไซต์อนาจารที่มีเส้นเชื่อมขี้ออกไปยังเว็บไซต์ต่างๆ เป็นจำนวนมาก และมีความสัมพันธ์โดยมีการสร้างเส้นเชื่อมระหว่าง seed set ด้วยกัน จะทำให้ LFPD สามารถตรวจจับเว็บไซต์อนาจารได้เป็นจำนวนมากกว่าเมื่อเปรียบเทียบในรอบการทำงานที่เท่ากัน

สรุปและข้อเสนอแนะ

สรุป

ในงานวิจัยนี้ ผู้วิจัยได้ออกแบบ Link Farm Based Pornographic Web Detection Algorithm หรือ LFPD ซึ่งเป็นอัลกอริทึมสำหรับตรวจจับเว็บไซต์อนาจารที่ทำงานแบบกึ่งอัตโนมัติ โดยมีจุดเด่นอยู่ที่ LFPD ต้องการความช่วยเหลือจากผู้เชี่ยวชาญในการกำหนด seed set สำหรับเริ่มต้นการทำงาน LFPD มีการทำงานแบ่งได้เป็นสามขั้นตอน ได้แก่ การสร้างกลุ่มเว็บไซต์อนาจารเริ่มต้น การรวบรวมข้อมูลเว็บ และการตรวจจับเว็บไซต์อนาจาร ในการตรวจจับเว็บไซต์อนาจารนั้น LFPD จะใช้การตรวจสอบโครงสร้างเส้นเชื่อมของกลุ่มเว็บไซต์อนาจาร เพื่อหาเส้นเชื่อมชี้ไปและกลับ (common link) ระหว่างเว็บไซต์ที่ถูกตรวจสอบกับเว็บไซต์อนาจาร หากเว็บไซต์ที่ถูกตรวจสอบมีจำนวน common link มากกว่าขีดเริ่มเปลี่ยน T ที่กำหนดเว็บไซต์นั้นจะถูกระบุให้เป็นเว็บไซต์อนาจารและนำเข้าสู่ seed set จากนั้น LFPD ก็จะทำงานในรอบต่อไป

ในการออกแบบ LFPD เราได้ศึกษาความสัมพันธ์ระหว่างโครงสร้างเส้นเชื่อมของเว็บไซต์อนาจารกับเว็บไซต์ทั่วไปโดยการทบทวนผลการศึกษาวิจัยที่ผ่านมาประกอบกับการสังเกตเพิ่มเติม พบว่าเราสามารถกำหนดสมมุติฐานของความสัมพันธ์ระหว่างโครงสร้างเส้นเชื่อมของเว็บไซต์อนาจารกับเว็บไซต์ทั่วไปได้สามประการ ได้แก่ โครงสร้างความสัมพันธ์ระหว่างกลุ่มเว็บไซต์อนาจารมีลักษณะเป็นลิงก์ฟาร์มโดยมีเส้นเชื่อมระหว่างกันอย่างหนาแน่น, เว็บไซต์อนาจารบางแห่งมีเส้นเชื่อมชี้ไปยังเว็บไซต์ทั่วไป และปกติแล้วเว็บไซต์ทั่วไปจะไม่สร้างเส้นเชื่อมชี้กลับมายังเว็บไซต์อนาจาร ซึ่งสมมุติฐานดังกล่าวได้ใช้เป็นหลักในการกำหนดกฎเกณฑ์การทำงานของ LFPD

ในการทดลองเรากำหนดเว็บไซต์สำหรับเริ่มต้นการทำงานหรือ seed set จำนวน 8 เว็บไซต์ เพื่อให้เพียงพอที่จะทดสอบสมมุติฐานการทดลองที่กำหนดขึ้น ทั้งนี้คุณลักษณะของเว็บไซต์อนาจารที่เหมาะที่จะเป็น seed set นั้น ควรเป็นเว็บไซต์อนาจารที่มีเส้นเชื่อมชี้ออกไปยังเว็บไซต์อนาจารอื่นๆ เป็นจำนวนมาก และควรมีความสัมพันธ์โดยมีการสร้างเส้นเชื่อมชี้ไปและกลับระหว่างเว็บไซต์ที่เป็น seed set ด้วยกัน

ผลการทดลองพบว่า LFPD สามารถตรวจจับเว็บไซต์อนาจารได้ตั้งแต่ 18,725 ถึง 44,123 เว็บไซต์ จากการทำงานจำนวน 3 รอบ ทั้งนี้จำนวนเว็บไซต์ที่ตรวจจับได้ขึ้นอยู่กับค่า T ที่ใช้ หาก T

มีค่ามากจำนวนเว็บไซต์ที่ตรวจจับได้ก็จะมีจำนวนลดลงตามลำดับ เมื่อพิจารณาขีดความสามารถของ LFPD ที่ให้ผลการตรวจจับเว็บไซต์อนาจารที่ความแม่นยำสูงกว่าร้อยละ 90 พบว่าเราสามารถตรวจจับเว็บไซต์อนาจารได้ 18,725 เว็บไซต์ เมื่อใช้ T 4 และ LFPD ทำงานครบ 3 รอบ

เนื่องจากค่า T มีผลต่อความแม่นยำและอัตราการตรวจจับ ดังนั้นในการใช้งาน LFPD เราจึงควรพิจารณาเลือกใช้ค่า T ที่เหมาะสม เช่น เลือกใช้ T ที่มีค่าต่ำหากเรารายชื่อเว็บไซต์อนาจารโดยไม่ต้องการความแม่นยำมากนัก เนื่องจากเราอาจมีระบบงานอื่นมาช่วยในการตรวจสอบและคัดกรองอีกชั้นหนึ่ง หรือเลือกใช้ T ที่มีค่าสูง หากเราต้องการความแม่นยำในการตรวจจับเว็บไซต์อนาจารเพื่อยืนยันรายชื่อเว็บไซต์อนาจารให้กับหน่วยงานหรือระบบงานอื่น เป็นต้น

ในการรวบรวมข้อมูลเว็บ เนื่องจากวัตถุประสงค์ของ LFPD คือการตรวจจับเว็บไซต์อนาจาร เราจึงต้องการรวบรวมข้อมูลเฉพาะเว็บไซต์อนาจารโดยไม่ต้องการรวบรวมข้อมูลเว็บไซต์ทั่วไป ดังนั้นเราจึงใช้ SPS เพื่อจำกัดการรวบรวมข้อมูลเว็บจากเว็บไซต์ที่มีแนวโน้มที่จะเป็นเว็บไซต์อนาจารโดยเป็นเว็บไซต์ที่ได้รับเส้นเชื่อมที่ชี้มาจากเว็บไซต์อนาจารที่ถูกตรวจจับแล้วอย่างน้อย 1 เส้น การใช้ SPS ดังกล่าวทำให้อัตราการตรวจจับของ LFPD มีค่าสูงขึ้น โดยเฉพาะเมื่อ LFPD ทำงานในจำนวนรอบที่มากขึ้นและเว็บไซต์ที่รวบรวมได้เริ่มมีเส้นเชื่อมชี้ไปยังเว็บไซต์ทั่วไปมากขึ้น

อย่างไรก็ตามจากผลการทดลองทำให้เราพบว่ามีการตรวจจับเว็บไซต์อนาจารผิดพลาด ซึ่งเป็นจุดอ่อนที่สำคัญของ LFPD ข้อผิดพลาดดังกล่าวเกิดขึ้นเนื่องจากมีเว็บไซต์ทั่วไปที่มีการสร้างความสัมพันธ์กับเว็บไซต์อนาจารอย่างจงใจ โดยการสร้างเส้นเชื่อมชี้ไปและกลับระหว่างกัน ปรากฏการณ์ดังกล่าวไม่สอดคล้องกับสมมุติฐานที่คาดการณ์ไว้ ส่งผลให้การทำงานของ LFPD ในรอบที่ 3 มีความแม่นยำลดลงจนมีค่าต่ำกว่าร้อยละ 90

นอกจากนี้ในการทดลองเรายังพบปัญหาต่างๆ เกี่ยวกับคุณลักษณะของเว็บไซต์อนาจารและข้อมูลที่ใช้ในการตรวจสอบเว็บไซต์อนาจาร ได้แก่ เว็บไซต์อนาจารบางเว็บไม่ปรากฏข้อมูลในรายชื่อเว็บไซต์อนาจารที่ได้รับจากเว็บไซต์ URLBlacklist.com เว็บไซต์อนาจารบางเว็บเป็นเว็บไซต์ที่หมดอายุ และเว็บไซต์อนาจารบางเว็บเป็นเว็บไซต์ที่ทดลองสร้างขึ้นมา โดยเว็บไซต์ดังกล่าวไม่พบว่ามีเส้นเชื่อมชี้เข้าหรือออกจากเว็บไซต์แต่อย่างใด ทั้งนี้ปัญหาดังกล่าวส่งผลต่อความสามารถและความถูกต้องในการตรวจจับเว็บไซต์อนาจารของ LFPD

ข้อเสนอแนะ

จากผลการทดลองเราพบจุดอ่อนของ LFPD หลายประการ การพัฒนา LFPD ต่อไปนั้น เราจะต้องปรับปรุงการทำงานของ LFPD ในด้านต่างๆ ประการแรกได้แก่การกำหนด seed set โดยผู้เชี่ยวชาญซึ่งมีจุดอ่อนในเรื่องความยากที่จะกำหนด seed set ให้ครอบคลุมเว็บไซต์อนาจารกลุ่มต่างๆ ได้ทั้งหมด เนื่องจากหากปรากฏกลุ่มเว็บไซต์อนาจารที่มิได้เชื่อมโยงระหว่างกันผ่านกลุ่มเว็บไซต์อนาจารหรือมีลักษณะเป็น disconnected graph และผู้เชี่ยวชาญไม่สามารถหรือมิได้กำหนด seed set ที่อยู่ในกลุ่มเว็บไซต์อนาจารที่มีลักษณะดังกล่าว ก็จะทำให้ไม่สามารถตรวจจับเว็บไซต์อนาจารในกลุ่มนั้นๆ ได้ ประการที่สองได้แก่การเพิ่มขีดความสามารถในการตรวจจับให้ถูกต้องมากยิ่งขึ้น ซึ่งเราต้องปรับสมมุติฐานที่ผิดพลาดให้สอดคล้องกับข้อเท็จจริงที่ว่ามีเว็บไซต์ทั่วไปบางแห่งสร้างเส้นเชื่อมชี้ไปและกลับร่วมกับเว็บไซต์อนาจาร ทั้งนี้การเพิ่มขีดความสามารถในการตรวจจับดังกล่าวอาจกระทำได้โดยใช้การตรวจสอบคำหรือการตรวจสอบภาพที่ปรากฏในแต่ละเว็บเพจ เพื่อระบุการเป็นเว็บไซต์อนาจารเพิ่มเติมจากการใช้โครงสร้างเส้นเชื่อมเพียงอย่างเดียว และเรายังอาจปรับปรุงเพิ่มเติมหลักเกณฑ์การคัดเลือกเว็บไซต์ที่ได้รับการตรวจจับใหม่ที่จะนำเข้าไปเพิ่มไว้ใน seed set สำหรับการทำงานในรอบถัดไป เพื่อเพิ่มความแม่นยำและอัตราการตรวจจับของ LFPD ให้สูงยิ่งขึ้น

นอกจากนี้ปัญหาที่พบในการทดลอง ได้แก่ ความถูกต้องข้อมูลที่ใช้ในการตรวจสอบยืนยันผลการทำงานของ LFPD ก็ควรจะต้องได้รับการปรับปรุง โดยอาจใช้ข้อมูลจากแหล่งข้อมูลอื่นเพิ่มเติม รวมทั้งการตรวจสอบเว็บไซต์อนาจารที่ LFPD ตรวจจับได้ แต่ไม่ปรากฏในรายชื่อเว็บไซต์อนาจารที่ใช้ตรวจสอบยืนยัน ซึ่งอาจต้องใช้ทีมผู้เชี่ยวชาญหรือการตรวจสอบด้วยวิธีการอื่นๆ เพื่อให้การตรวจสอบยืนยันผลการทดลองมีความสมบูรณ์และถูกต้องยิ่งขึ้นต่อไป

เอกสารและสิ่งอ้างอิง

พิรงรอง รามสูต รณะนันทน์ และนิธิมา คณานิธินันท์. 2547. **รายงานวิจัยฉบับสมบูรณ์เรื่อง การกำกับดูแลเนื้อหาอินเทอร์เน็ต**. สำนักงานกองทุนสนับสนุนการวิจัย

Bharat, K., B.W. Chang, M. Henzinger and M. Ruhl. 2001. Who links to whom: mining linkage between web sites, pp. 51-58. **In Proceedings of the IEEE International Conference on Data Mining**, **November 2001**, San Jose, California

Björneborn, L. and P. Ingwersen. 2004. Towards a basic framework of webometrics. **Journal of the American Society for Information Science and Technology**, 55(14), 1216-1227.

Chan, Y., R. Harvey and D. Smith. 1999. Building systems to block pornography. **In Challenge of Image Retrieval, 25-26 February 1999**, Newcastle.

Gyongyi, Z. and H. Garcia-Molina. 2005. Web spam taxonomy, **In Proceedings of First International Workshop on Adversarial Information Retrieval on the Web**, Chiba, Japan

_____, _____ and J. Pedersean. 2004. Combating web spam with TrustRank, pp. 576-587. **In Proceedings of the Thirtieth International Conference on Very Large Database**. Very Large Database Endowment, Toronto. Canada

Internet Content Rating Association. 2008. **About ICRA.** Family Online Safety Institute. Available Source: http://www.fosi.org/icra/, March 23, 2008

Internet World Stats. 2008. **Internet World Stats Blog for 2008.** Internet World Stats. Available Source: http://www. internetworldstats.com/, May 21, 2008.

Lee, P.Y., S.C. Hui and A. C. M. Fong. 2002. Neural Networks for Web Content Filtering. **IEEE Intelligent Systems,** (17): 48-57.

Lin, Y.C., H. W. Tseng and C. S. Fuh. 2003. Pornography Detection Using Support Vector Machine. **Sixteenth IPPR Conference on Computer Vision, Graphics and Image Processing,** Kinmen, Republic of China.

Page, L., S. Brin, R. Motwani and T. Winograd. 1998. **The PageRank Citation Ranking : Bringing Order to the Web**. Stanford Digital Library Technologies Project, Stanford University.

Resnick, P. and J. Miller. 2002. **PICS: Internet access controls without censorship.** W3C. Available Source: http://www.w3.rot/PICS/iacwc.htm, March 23, 2008

URLBlacklist.com. 2008. **URLBlacklist.com.** URLBlacklist.com. Available Source: http://urlblacklist.com/, March 23, 2008

Wu, B. and B. D. Davison. 2005. Identifying Link Farm Spam Pages. **In Proceedings of the Fourteenth International World Wide Web Conference**, Chiba, Japan.

_____, V. Goel and _____ 2006. Topical TrustRank: Using Topicality to Combat Web Spam. **In Proceedings of the Fifteenth International World Wide Web Conference**, Edinburgh, Scotland

ภาคผนวก

ภาคผนวก ก

รายละเอียดผลการทดลอง

ตารางผนวกที่ ก1 ผลการทำงานของ LFPD เมื่อไม่ใช้ SPS

รายการ	ผลการทำงาน		
	รอบที่ 1	**รอบที่ 2**	**รอบที่ 3**
T 1			
เว็บไซต์ที่ตรวจจับได้ทั้งหมด	445	12,054	65,077
เว็บไซต์ที่ตรวจจับได้และตรงกับ Blacklist	430	11,097	50,036
seedset + SPS ที่ตรงกับ Blacklist	1,229	21,521	110,546
seedset + SPS	1,360	26,064	190,639
ความแม่นยำ(Precision)(ร้อยละ)	96.63	92.06	76.89
อัตราการตรวจจับ (Detection Rate) (ร้อยละ)	31.62	42.58	26.25
T 2			
เว็บไซต์ที่ตรวจจับได้ทั้งหมด	381	9,101	45,503
เว็บไซต์ที่ตรวจจับได้และตรงกับ Blacklist	368	8,497	38,156
seedset + SPS ที่ตรงกับ Blacklist	1,229	21,521	110,546
seedset + SPS	1,360	26,064	190,639
ความแม่นยำ(Precision)(ร้อยละ)	96.59	93.36	83.85
อัตราการตรวจจับ (Detection Rate) (ร้อยละ)	27.06	32.60	20.01
T 3			
เว็บไซต์ที่ตรวจจับได้ทั้งหมด	325	8,056	37,735
เว็บไซต์ที่ตรวจจับได้และตรงกับ Blacklist	313	7,545	32,731
seedset + SPS ที่ตรงกับ Blacklist	1,229	21,521	110,546
seedset + SPS	1,360	26,064	190,639
ความแม่นยำ(Precision)(ร้อยละ)	96.31	93.66	86.74
อัตราการตรวจจับ (Detection Rate) (ร้อยละ)	23.01	28.95	17.17

ตารางผนวกที่ ก1 (ต่อ)

รายการ	ผลการทำงาน		
	รอบที่ 1	**รอบที่ 2**	**รอบที่ 3**
T 4			
เว็บไซต์ที่ตรวจจับได้ทั้งหมด	248	7,285	32,005
เว็บไซต์ที่ตรวจจับได้และตรงกับ Blacklist	237	6,840	28,327
seedset + SPS ที่ตรงกับ Blacklist	1,229	21,521	110,546
seedset + SPS	1,360	26,064	190,639
ความแม่นยำ(Precision)(ร้อยละ)	95.56	93.89	88.51
อัตราการตรวจจับ (Detection Rate) (ร้อยละ)	17.43	26.24	14.86

ภาคผนวก ข

ผลงานตีพิมพ์

1. อัลกอริทึมในการตรวจจับเว็บอนาจารด้วยลิงก์ฟาร์ม

Link Farm Based Pornographic Web Detection Algorithm

(CIT 2009), January 14-17, 2009

อัลกอริทึมในการตรวจจับเว็บอนาจารด้วยลิงก์ฟาร์ม
Link Farm Based Pornographic Web Detection Algorithm

สุเอซ อมาตยกุล
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
มหาวิทยาลัยเกษตรศาสตร์
g4864207@ku.ac.th

ผศ.ดร.สุขุมาล กิติสิน
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
มหาวิทยาลัยเกษตรศาสตร์
Sukumal.i@ku.ac.th

Abstract

At present, there are a lot of pornographic websites. URL blocking is the one technique of web filtering to be protected the access to them. The major problem of this technique is URL list implement spends a large team of specialists. So, This study presents the semi-automatic technique to detect and identify the pornographic website, called "Link Farm Based Pornographic Web Detection Algorithm (LFPD)". LFPD detects relation of link structure of pornographic websites. There are three steps; to identify the seed set of pornographic website by specialist, to collect of website data and to examine and identify pornographic website. The experimental results show that examine and identify pornographic web site of 400 times of seed set with the accuracy more than 90%

Keywords: Pornographic Website, Web filtering, Web detection, Web structure

บทคัดย่อ

ปัจจุบันมีเว็บไซต์ที่เผยแพร่เนื้อหาอนาจารจำนวนมาก การกรองเว็บโดยใช้บัญชีรายชื่อเว็บไซต์ (URL blocking) เป็นเทคนิคหนึ่งในการป้องกันการเข้าถึงเว็บไซต์ดังกล่าว การจัดทำบัญชีรายชื่อเว็บไซต์อนาจารโดยผู้เชี่ยวชาญเป็นงานที่สิ้นเปลืองเวลาและแรงงาน ดังนั้นในงานวิจัยนี้จึงเสนออัลกอริทึมสำหรับตรวจสอบและบ่งชี้เว็บไซต์อนาจารแบบกึ่งอัตโนมัติ เรียกว่า Link Farm Based Pornographic Web Detection Algorithm (LFPD) โดยเป็นการตรวจหาความสัมพันธ์ของโครงสร้างเส้นเชื่อมระหว่างกลุ่มเว็บไซต์อนาจาร LFPD มีการทำงานสามขั้น ได้แก่ การกำหนดกลุ่มเว็บไซต์อนาจารเริ่มต้นโดยผู้เชี่ยวชาญ การรวบรวมข้อมูลเว็บไซต์ และการตรวจสอบและบ่งชี้เว็บไซต์อนาจาร ผลการทดลองพบว่า LFPD สามารถตรวจจับเว็บไซต์อนาจารได้ไม่น้อยกว่า 400 เท่าของกลุ่มเว็บไซต์อนาจารเริ่มต้น ที่ค่าความแม่นยำสูงกว่าร้อยละ 90

คำสำคัญ: เว็บไซต์อนาจาร, การกรองเว็บ, การตรวจจับเว็บ, โครงสร้างเว็บ

1. บทนำ

ปัจจุบันมีเว็บไซต์ที่มีเนื้อหาอนาจารจำนวนมาก ดังนั้นจึงมีการป้องกันการเข้าถึงเว็บไซต์อนาจาร โดยตรวจจับข้อความหรือภาพอนาจาร หรือปิดกั้นโดยใช้บัญชีรายชื่อเว็บไซต์อนาจาร ปัญหาในการจัดทำบัญชีรายชื่อเว็บไซต์ ได้แก่ การรวบรวมและตรวจสอบรายชื่อ เนื่องจากเว็บไซต์อนาจารมีจำนวนมากและเพิ่มขึ้นทุกวัน ทำให้การรวบรวมและตรวจสอบรายชื่อเว็บไซต์โดยผู้เชี่ยวชาญเป็นงานที่สิ้นเปลืองแรงงานและเวลาเป็นอย่างมาก

จากการศึกษาโครงสร้างเส้นเชื่อมกลุ่มเว็บไซต์อนาจารพบว่ามีการสร้างเส้นเชื่อมระหว่างเว็บไซต์ภายในกลุ่มเป็นจำนวนมาก ทำให้โครงสร้างเส้นเชื่อมของกลุ่มเว็บไซต์อนาจารคล้ายกับลิงก์ฟาร์ม (link

farm) ซึ่งมีผู้ศึกษาวิจัยเกี่ยวกับการตรวจหาลิงก์ฟาร์มที่เกิดจากการสแปมเว็บโดยใช้โครงสร้างเส้นเชื่อม

ในงานวิจัยนี้จะนำเทคนิคการตรวจจับสแปมเพจข้างต้นมาประยุกต์ใช้ เพื่อตรวจสอบและบ่งชี้เว็บไซต์อนาจารแบบกึ่งอัตโนมัติ ซึ่งจะสามารถลดระยะเวลาและแรงงานเมื่อเปรียบเทียบกับการตรวจหาด้วยผู้เชี่ยวชาญเพียงอย่างเดียว

2. หลักการที่เกี่ยวข้อง

2.1 ระบบการกรองเว็บ (web filtering)

ระบบกรองเว็บในปัจจุบันแบ่งได้เป็น 4 ลักษณะ ได้แก่

2.1.1 Keyword blocking เป็นการใช้ชุดคำสำคัญ (keyword) ที่จัดทำขึ้นเพื่อระบุเว็บเพจที่ไม่เหมาะสม [1]

2.1.2 Image blocking เป็นการตรวจจับสีของผิวหนังมนุษย์ที่ปรากฏ (skin detection) ตามเกณฑ์ที่กำหนด [2]

2.1.3 Rating System เป็นการกลั่นกรองเว็บโดยให้ผู้พัฒนาเว็บหรือองค์กรอิสระพิจารณาเนื้อหา และติดสัญลักษณ์ (label) ระบุประเภทของเนื้อหาไว้ที่เว็บเพจ [3]

2.1.4 URL blocking เป็นการปิดกั้นโดยใช้บัญชีรายชื่อเว็บที่ไม่ปลอดภัย (blacklists) เทคนิคการกรองเว็บนี้มีปัญหาสำคัญเนื่องจากเว็บไซต์อนาจารมีจำนวนมากและเพิ่มขึ้นทุกวัน จึงต้องปรับปรุงรายชื่อเว็บไซต์อย่างต่อเนื่อง การใช้ผู้เชี่ยวชาญรวบรวมและตรวจสอบรายชื่อเว็บไซต์อนาจารเพียงอย่างเดียวจึงสิ้นเปลืองเวลาและแรงงานเป็นอย่างมาก [1]

2.2 โครงสร้างเว็บ (web structure)

โครงสร้างเว็บเปรียบได้กับกราฟแบบมีทิศทาง (direct graph) เว็บโหนด หมายถึง เว็บเพจ เว็บไดเร็คทอรี่ หรือเว็บไซต์ อย่างใดอย่างหนึ่ง เทียบได้กับโหนด (nodes) ของกราฟ และไฮเปอร์ลิงก์ (hyperlinks) หรือเส้นเชื่อมระหว่างเว็บโหนด เทียบได้กับเส้นเชื่อม (egdes) ของกราฟ [4]

ในปี 1998 Larry Page และคณะเปรียบเทียบผลการจัดอันดับโดย PageRank กับข้อมูลการใช้งานเว็บ (web usage) พบว่ามีการใช้งานเว็บไซต์อนาจารจำนวนมาก แต่ PageRank ของกลุ่มเว็บไซต์อนาจารกลับมีค่าไม่สูงนัก จึงสรุปว่า “ในขณะที่คนทั่วไปมักเข้าใช้เว็บไซต์อนาจาร แต่ก็ไม่ต้องการสร้างเส้นเชื่อมที่ชี้จากเว็บไซต์ของพวกเขาไปยังเว็บไซต์อนาจาร” [5]

2.3 การสแปมเว็บ (web spam)

การสแปมเว็บเป็นความพยายามสร้างหรือดัดแปลงเว็บเพจ เพื่อให้เครื่องมือค้นหาข้อมูลบนเว็บ (search engine) ได้รับข้อมูลนอกเหนือจากที่ควรส่งผลให้เว็บเพจเป้าหมายได้รับการจัดอันดับสูงกว่าปกติ [6] ทั้งนี้การสแปมเว็บเพื่อเพิ่มผลการจัดอันดับมักปรากฏโดยเฉพาะในกลุ่มของเว็บไซต์อนาจาร [7]

2.4 ลิงก์ฟาร์ม (link farm)

ลิงก์ฟาร์ม หมายถึง กลุ่มของเว็บเพจที่มีเส้นเชื่อมระหว่างกันอย่างหนาแน่น [8] ลิงก์ฟาร์มที่สร้างขึ้นเพื่อเพิ่มผลการจัดอันดับเว็บ เรียกว่า สแปมฟาร์ม (spam farm) [6] เพื่อให้การจัดอันดับเว็บเที่ยงตรง จึงมีการตรวจหาสแปมฟาร์ม เพื่อนำข้อมูลไปปรับปรุงผลการจัดอันดับเว็บต่อไป

2.4.1 TrustRank [9] Gyongyi และคณะเสนออัลกอริทึม TrustRank ซึ่งใช้แยกเว็บเพจดีออกจากสแปมเพจแบบกึ่งอัตโนมัติ โดยมีแนวความคิดว่า “เว็บเพจดีมักชี้ไปยังเว็บเพจดีด้วยกัน และน้อยครั้งที่จะชี้ไปยังเว็บเพจที่ไม่ดี” การทำงานของ TrustRank จะใช้ผู้เชี่ยวชาญเลือกเว็บเพจดีจำนวนหนึ่งจากเว็บเพจทั้งหมดเพื่อใช้เป็นข้อมูลเริ่มต้น การเลือกเว็บเพจเริ่มต้นโดยผู้เชี่ยวชาญทำให้มั่นใจได้ว่าเว็บเพจที่ถูกเลือกเป็นเว็บเพจดี อย่างไรก็ตามเป็นการยากที่จะ

เลือกกลุ่มของเว็บเพจดีให้เพียงพอที่จะเป็นตัวแทนของเว็บไซต์แต่ละกลุ่ม [10]

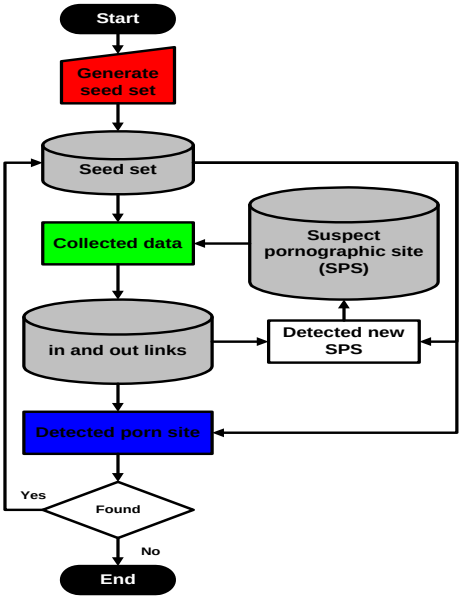
2.4.2 Wu และ Davison [8] เสนอการตรวจจับสแปมเพจ (spam pages) ภายในลิงก์ฟาร์มแบบอัตโนมัติ โดยมีการทำงานสามขั้นตอน ได้แก่

- การสร้าง seed set เป็นการตรวจหากลุ่มของสแปมเพจเริ่มต้น โดยตรวจหาเว็บเพจที่มีเส้นเชื่อมชี้ไปกลับ (reciprocal links) ซึ่งในงานวิจัยนี้เรียกว่า common link ระหว่างกัน และเรียกคู่ของเว็บเพจดังกล่าวว่า common nodes กำหนดขีดเริ่มเปลี่ยน (threshold) T_{io} เพื่อใช้เป็นค่าต่ำสุดในการตัดสินว่าจะต้องมี common nodes จำนวนเท่าใดเว็บเพจจึงจะถือเป็น seed set ของสแปมเพจภายในลิงก์ฟาร์ม
- การขยายการตรวจจับ (Expansion step) สแปมเพจจะมีเส้นเชื่อมชี้ไปยังสแปมเพจด้วยกัน หากเว็บเพจใดมีเส้นเชื่อมชี้ไปยังสแปมเพจมากกว่าค่าที่กำหนดจะถือว่าเว็บเพจดังกล่าวเป็นสแปมเพจ โดยกำหนดขีดเริ่มเปลี่ยน T_{pp} (Threshold ParentPenalty) เพื่อใช้ในการตัดสินการเป็นสแปมเพจ
- การปรับการจัดอันดับ เป็นการปรับปรุงลการจัดอันดับเว็บเพจให้ถูกต้อง

เทคนิคในการตรวจจับสแปมเพจของ Wu และคณะมีจุดเด่นที่ทำงานแบบอัตโนมัติ อย่างไรก็ตามไม่อาจเชื่อมั่นได้ว่าเว็บเพจที่ถูกระบุในขั้นการตรวจหา seed set จะเป็นสแปมเพจ

3. การตรวจจับและยืนยันเว็บไซต์อนาจาร

เนื่องจากโครงสร้างเส้นเชื่อมของกลุ่มเว็บไซต์อนาจารมีลักษณะคล้ายคลึงกันกับลิงก์ฟาร์ม เราจึงพัฒนาเทคนิคการตรวจจับเว็บไซต์อนาจาร เรียกว่า Link Farm Based Pornographic Web Detection Algorithm (LFPD) โดยนำการตรวจจับสแปมฟาร์มมาประยุกต์ใช้ในการตรวจจับเว็บไซต์อนาจาร ด้วยการผสานการกำหนดเว็บเพจเริ่มต้นด้วยผู้เชี่ยวชาญของ TrustRank [9] และการหาตรวจหา seed set ของ Wu และ Davison [8] เข้าด้วยกัน

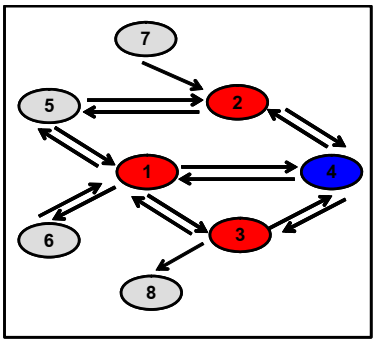


รูปที่ 1 แสดงขั้นตอนการทำงาน

3.1 ขั้นตอนวิธี

LFPD ทำงานในลักษณะเวียนเกิด (recursive) แบ่งการทำงานเป็นสามขั้นตอน ตามรูปที่ 1 ได้แก่

- Generate seed set เป็นการสร้าง seed set ของเว็บไซต์อนาจารโดยผู้เชี่ยวชาญ
- Collected data เป็นการรวบรวมข้อมูลเส้นเชื่อมที่ชี้เข้าและออกจาก seed set และเว็บไซต์ต้องสงสัย
- Detected porn site เป็นการตรวจจับและบ่งชี้เว็บไซต์อนาจาร หากพบเว็บไซต์อนาจารใหม่ก็จะกลับไปรวบรวมข้อมูลเว็บเพิ่มเติมจนกว่าจะครบจำนวนรอบที่กำหนดหรือไม่พบเว็บไซต์อนาจาร



รูปที่ 2 แสดงภาพจำลองเว็บไซต์จำนวน 7 เว็บไซต์

3.2 การบ่งชี้เว็บไซต์อนาจาร

การบ่งชี้เว็บไซต์อนาจารกระทำโดยการหาคู่ของเว็บไซต์ (common nodes) ที่มีเส้นเชื่อมชี้ไปและกลับ(common link) ระหว่างเว็บไซต์ที่ตรวจสอบกับเว็บไซต์อนาจาร จากรูปที่ 2 หากเว็บไซต์ 1, 2 และ 3 เป็นเว็บไซต์อนาจารที่ได้รับการยืนยัน และเว็บไซต์ 4 เป็นเว็บไซต์ที่ถูกตรวจสอบ เห็นได้ว่าเว็บไซต์ 4 มี common link หรือมีเส้นเชื่อมชี้เข้าและออกร่วมกับเว็บไซต์ 1, 2 และ 3 หรือกล่าวได้ว่าเป็น common nodes กับเว็บไซต์ 1, 2 และ 3 หากเรากำหนดขีดต่ำสุด(Threshold, TH) สำหรับตัดสินว่าเว็บไซต์ใดจะเป็นเว็บไซต์อนาจารเท่ากับสาม ดังนั้นเว็บไซต์ 4 จะถูกพิจารณาว่าเป็นเว็บไซต์อนาจาร สำหรับเว็บไซต์ 5 และเว็บไซต์ 6 ไม่เป็นเว็บไซต์อนาจาร เนื่องจากมี common nodes ร่วมกันเว็บไซต์อนาจารน้อยกว่าสาม ส่วนเว็บไซต์ 7 และ 8 ก็ไม่เป็นเว็บไซต์อนาจารเช่นกัน เนื่องจากไม่มี common link ร่วมกับเว็บไซต์อนาจาร

3.3 เว็บไซต์ต้องสงสัย (suspect pornographic site, SPS)

เนื่องจากเว็บไซต์อนาจารอาจมีเส้นเชื่อมชี้ไปยังเว็บไซต์ทั่วไป หากเรารวบรวมเว็บไซต์ทั้งหมดที่พบก็อาจรวบรวมข้อมูลเว็บไซต์อนาจารและเว็บไซต์ทั่วไปไปพร้อมกัน ทำให้สูญเสียประสิทธิภาพในการรวบรวมข้อมูล ดังนั้นเราจึงจะรวบรวมข้อมูลจากเว็บไซต์ที่มีแนวโน้มว่าน่าจะเป็นเว็บไซต์อนาจาร โดยถือว่าเว็บไซต์ได้รับเส้นเชื่อมชี้ออกมาจากเว็บไซต์อนาจารเป็นเว็บไซต์ที่มีแนวโน้มที่จะเป็นเว็บไซต์อนาจาร หรือเรียกว่าเว็บไซต์ต้องสงสัย (suspect pornographic site, SPS) ซึ่งในการทดลองนี้ SPS ได้แก่ เว็บไซต์ที่ได้รับเส้นเชื่อมที่ชี้มาจากเว็บไซต์อนาจารอย่างน้อย 1 เส้น

4. วิธีการทดลอง

4.1 เครื่องมือและการรวบรวมข้อมูล

เราพัฒนาโปรแกรมรวบรวมข้อมูลเว็บ (web crawler) และ LFPD ด้วยภาษา JAVA โดยใช้ NetBeans IDE 6.1 และเก็บข้อมูลเว็บที่รวบรวมได้ในฐานข้อมูล MySQL Server 5.0

เนื่องจากเราต้องการเปรียบเทียบการทำงานของ LFPD ที่ค่า Threshold(TH) ต่างกัน และต้องการทดสอบการขยายการตรวจจับโดยไม่ต้องรวบรวมข้อมูลเว็บมากเกินความจำเป็น ในการทดลองนี้จึงใช้ seed set สำหรับเริ่มต้นการทำงานของ LFPD จำนวน 8 เว็บไซต์

4.2 การประเมินผลการทดลอง (Evaluation)

การประเมินผลการทดลองกระทำโดยการเปรียบเทียบความแม่นยำ (precision) ความระลึก (recall) และประสิทธิภาพ (efficiency) ของ LFPD ที่ TH ที่แตกต่างกัน โดยข้อมูลที่ใช้ยืนยันผลการทำงานของ LFPD นั้น นำมาจากรายชื่อเว็บไซต์อนาจารที่เผยแพร่ใน URLBlasklist.com [11]

5. ชุดข้อมูลและผลการทดลอง

5.1 ชุดข้อมูล

จาก seed set เริ่มต้น จำนวน 8 เว็บไซต์ ทำการเก็บข้อมูลเว็บในห้วงกันยายน - ตุลาคม 2551 ข้อมูลที่ได้ประกอบด้วย 3,550,803 URLs จาก 190,733

เว็บไซต์ และเราบ่งชี้เว็บไซต์อนาจารได้โดยใช้ค่า TH ในห้วงตั้งแต่ 1 – 4

5.2 ผลการทดลอง

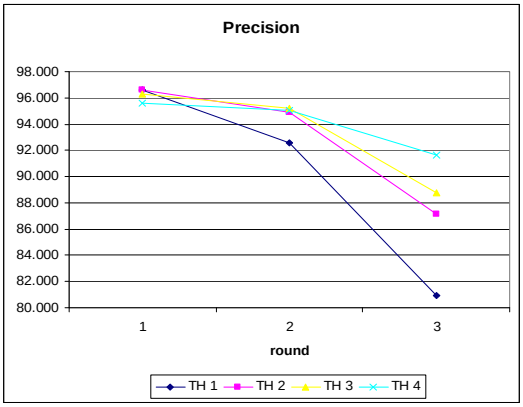
รายการ	Threshold (TH)			
	1	2	3	4
เว็บไซต์บ่งชี้	10,264	7,401	5,820	3,564
เว็บไซต์บ่งชี้และตรงกับ Blacklist	9,500	7,022	5,539	3,388
Seed set และ SPS ที่ตรงกับ Blacklist	19,259	18,422	16,433	11,110

ตารางที่ 1 แสดงผลการทดลองในรอบที่ 2

รายการ	Threshold (TH)			
	1	2	3	4
เว็บไซต์บ่งชี้	54,518	37,069	29,916	20,440
เว็บไซต์บ่งชี้และตรงกับ Blacklist	44,123	32,298	26,549	18,725
Seed set และ SPS ที่ตรงกับ Blacklist	94,329	84,786	77,794	58,746

ตารางที่ 2 แสดงผลการทดลองในรอบที่ 3

ผลการทดลองในตารางที่ 1 และ 2 แสดงให้เห็นว่า เมื่อ LFPD ทำงานไปครบ 3 รอบ เราสามารถตรวจสอบและบ่งชี้เว็บไซต์อนาจารได้ตั้งแต่ 18,725 จนถึง 44,123 เว็บไซต์ ขึ้นอยู่กับ TH หาก TH มีค่าเพิ่มขึ้น เว็บไซต์อนาจารที่ตรวจจับได้ในแต่ละรอบจะลดลงตามลำดับ



รูปที่ 3 กราฟแสดงค่าความแม่นยำ (Precision)

เนื่องจากเว็บไซต์อนาจารที่ถูกบ่งชี้ในรอบก่อนหน้าจะถูกนำกลับมาใช้เป็น seed set ในรอบถัดไป ประสิทธิภาพในแต่ละรอบการทำงานของ LFPD จึงขึ้นอยู่กับความถูกต้องของการบ่งชี้ในรอบก่อนหน้า ทำให้ LFPD เกิดข้อจำกัดเนื่องจากมีเว็บไซต์ที่บ่งชี้ผิดพลาดสะสมเพิ่มมากขึ้นในแต่ละรอบ โดยผลการทดลองในรอบที่ 3 ตามรูปที่ 3 เห็นได้ว่าความแม่นยำ (precision) ลดลงอย่างมาก เนื่องจากการบ่งชี้ผิดพลาดสะสมในรอบที่ 1 และ 2

อย่างไรก็ตามสรุปได้ว่า “เมื่อใช้ seed set ซึ่งเป็นเว็บไซต์อนาจารที่อยู่ภายในกลุ่มของเว็บไซต์อนาจารจำนวนหนึ่ง LFPD สามารถตรวจสอบและบ่งชี้เว็บไซต์อนาจารเพิ่มเติมได้ภายใต้การทำงานไม่น้อยกว่า 2 รอบ โดยมีความแม่นยำสูงกว่าร้อยละ 90 หรือมีอัตราการขยายการตรวจจับในห้วงประมาณ 400 – 1,000 เท่า”

เมื่อตรวจสอบข้อมูลเว็บไซต์อนาจารที่บ่งชี้ผิดพลาดพบว่า การบ่งชี้ผิดพลาดจะเพิ่มขึ้นตามจำนวนรอบการทำงานที่เพิ่มขึ้น และยังสัมพันธ์กับค่า TH ที่ต่ำลง เว็บไซต์ทั่วไปที่ถูกบ่งชี้เป็นเว็บไซต์อนาจารสามารถแยกเป็นกลุ่มที่สำคัญ เช่น กลุ่มดาวน์

โหลดภาพยนตร์, กลุ่มการบริการเงินสด , กลุ่มการพนัน, กลุ่มหาคู่, กลุ่มเกมส์ เป็นต้น

ดังนั้นสรุปได้อีกว่า "มีเว็บไซต์ที่มิใช่เว็บไซต์อนาจารซึ่งมีเส้นเชื่อมชี้มายังเว็บไซต์อนาจารโดยเจตนา"

6. สรุปและแนวทางการวิจัยในอนาคต

งานวิจัยนี้เป็นการนำเทคนิคในการตรวจหาสแปมเพจภายในลิงก์ฟาร์มโดยใช้โครงสร้างเส้นเชื่อมมาประยุกต์ใช้ในการตรวจสอบและบ่งชี้เว็บไซต์อนาจาร เรียกว่า LFPD ผลการทดลองพบว่า เมื่อติดตามความสัมพันธ์ของเส้นเชื่อมชี้ไปกลับ (common link) ระหว่าง seed set กับเว็บไซต์ที่มีความสัมพันธ์กับ seed set เราสามารถขยายการตรวจหากลุ่มของเว็บไซต์อนาจารเพิ่มเติมได้ไม่น้อยกว่า 400 เท่าของจำนวน seed set เริ่มต้น อย่างไรก็ตามเราพบว่ามีเว็บไซต์ทั่วไปบางประเภทที่มีการสร้างเส้นเชื่อมชี้ไปกลับกับกลุ่มของเว็บไซต์อนาจารโดยเจตนา ดังนั้นความแม่นยำของ LFPD จึงจำกัดอยู่ที่การทำงานไม่เกิน 2 รอบ ดังนั้นปัจจัยหลักในการปรับปรุงประสิทธิภาพของ LFPD จึงอยู่ที่การลดความผิดพลาดในการบ่งชี้ ซึ่งอาจเพิ่มเติมการบ่งชี้เว็บไซต์อนาจารในลักษณะอื่น เช่น การตรวจสอบคำหรือข้อความที่ปรากฏ ฯลฯ ซึ่งเชื่อว่าจะลดข้อจำกัดในการทำงานดังกล่าว

เอกสารอ้างอิง

[1] Lee, P.Y., S.C. Hui and A. C. M. Fong. 2002. Neural Networks for Web Content Filtering. Neural Networks IEEE INTELLIGENT SYSTEMS, September-October 2002. 48-57

[2] Chandrinos, K. V., I. Androutsopoulos, G. Paliouras and C. D. Spyropoulos. 2000. Automatic Web Rating: Filtering Obscene Content on the Web. The 4th European Conference on Research and Advanced Technology for Digital Libraries. 403 - 406.

[3] Resnick, P. and J. Miller. 2002. PICS: Internet access controls without censorship. W3C. Available Source: http://www.w3.rot/PICS/iacwc.htm

[4] Björneborn, L. and P. Ingwersen. 2004. Towards a basic framework of webometrics. Journal of the American Society for Information Science and Technology, 55(14), 1216-1227.

[5] Page, L., S. Brin, R. Motwani and T. Winograd. 1998. The PageRank Citation Ranking.Proceeding of 1998 International Web Conference. 161: 172.

[6] Gyongyi, Z. and H. Garcia-Molina. 2005. Web spam taxonomy, In Proceedings of 1st International Workshop on Adversarial Information Retrieval on the Web.

[7] Bharat, K., B.W. Chang, M. Henzinger and M. Ruhl. 2001. Who links to whom: mining linkage between web sites, In: Proceedings of the IEEE international conference on data mining, San Jose, November 2001, pp. 51-58.

[8] Wu, B. and B. D. Davison. 2005. Identifying Link Farm Spam Pages. WWW2005, Chiba, Japan

[9] Gyongyi, Z. and H. Garcia-Molina and J. Pedersean. 2004. Combating web spam with TrustRank, In Proceedings of the 30th VLDB Conference.

[10] Wu, B. V. Goel and B. D. Davison. 2006. Topical TrustRank: Using Topicality to Combat Web Spam. WWW2006, Edinburgh, Scotland

[11] URLBlacklist.com. 2008. URLBlacklist.com. URLBlacklist.com. Available Source: http://urlblacklist.com/

ประวัติการศึกษาและการทำงาน

ชื่อ –นามสกุล	พันเอก สุเอซ อมาตยกุล
วัน เดือน ปี ที่เกิด	8 เมษายน พ.ศ.2512
สถานที่เกิด	กรุงเทพมหานคร
ประวัติการศึกษา	วท.บ.(ทบ.) โรงเรียนนายร้อยพระจุลจอมเกล้า (พ.ศ.2533)
ประวัติการทำงาน	- หัวหน้าแผนกกรรมวิธีข้อมูล กองสถิติ กรมส่งกำลังบำรุงทหารบก - รองผู้อำนวยการกอง กรมส่งกำลังบำรุงทหารบก - หัวหน้ากอง กรมส่งกำลังบำรุงทหารบก
ผลงานตีพิมพ์	สุเอซ อมาตยกุล และสุขุมาล กิติสิน. 2552. อัลกอริทึมในการตรวจจับเว็บอนาจารด้วยลิงก์ฟาร์ม. การประชุมวิชาการทางคอมพิวเตอร์และเทคโนโลยีสารสนเทศ (CIT2009), มหาวิทยาลัยขอนแก่น วิทยาเขตหนองคาย.