การตรวจจับลิงก์ฟาร์มโดยใช้การอนุมานไวยากรณ์กราฟ

นายวุฒิชัย วงศ์สารสิน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2553

LINK FARM DETECTION USING GRAPH GRAMMAR INFERENCE

Mr.Wuttichai Wongsarasin

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2010

หัวข้อวิทยานิพนธ์	การตรวจจับลิงก์ฟาร์มโดยใช้การอนุมานไวยากรณ์กราฟ
โดย	นายวุฒิชัย วงศ์สารสิน
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ผู้ช่วยศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง)

..... กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.เยาวดี เต็มธนาภัทร์)

วุฒิชัย วงศ์สารสิน : การตรวจจับลิงก์ฟาร์มโดยใช้การอนุมานไวยากรณ์กราฟ. (LINK FARM DETECTION USING GRAPH GRAMMAR INFERENCE) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : ผู้ช่วยศาสตราจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์, อ. ที่ปรึกษาวิทยานิพนธ์ร่วม: ผู้ช่วยศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง, 53 หน้า.

ลิงก์ฟาร์ม คือกลุ่มของเว็บเพจที่ถูกสร้างขึ้นเพื่อเพิ่มคะแนนการจัดอันดับให้กับเว็บเพจเป้าหมาย ซึ่งจากการกระทำดังกล่าวทำให้หน้าเว็บเพจที่มีคะแนนการจัดอันดับถูกจัดอยู่ในลำดับต้นๆ ของผลการค้นคืนของระบบสืบค้น ดังนั้นงานวิจัยจำนวนหนึ่งได้ถูกคิดค้นขึ้นเพื่อตรวจจับลิงก์ฟาร์ม สำหรับในงานวิจัยนี้ นำเสนอการอนุมานไวยากรณ์กราฟลิงก์ฟาร์มจากข้อมูลโครงสร้างเว็บสแปมบนเว็บกราฟ ซึ่งในการอนุมานใช้การพิจารณารูปแบบลิงก์ที่อยู่รอบโฮสต์เป้าหมายด้วยอัลกอริทึมป้ายกำกับ โดยที่รูปแบบของลิงก์สามารถบ่งบอกถึงลักษณะเฉพาะของโฮสต์สแปมที่แตกต่างไปจากโฮสต์ปกติ โดยผลลัพธ์ที่ได้จากการอนุมานคือไวยากรณ์จากโครงสร้างลิงก์ฟาร์ม และ โครงสร้างที่แสดงถึงลักษณะเฉพาะของลิงก์ฟาร์มจากการวัดประสิทธิภาพโดยทำการเปรียบเทียบกับจากวิจัยที่เกี่ยวข้องพบว่า การตรวจจับลิงก์ฟาร์มโดยใช้การอนุมานไวยากรณ์กราฟสามารถตรวจจับลิงก์ฟาร์มได้อย่างมีประสิทธิภาพ โดยเฉพาะอย่างยิ่งเว็บสแปมที่มีคะแนนการจัดอันดับสูง ซึ่งเป็นเป้าหมายหลักในการกำจัดออกจากระบบสืบค้น

ภาควิชา...วิศวกรรมคอมพิวเตอร์... ลายมือชื่อนิสิต...วุฒิชัย วงศ์สารสิน...
สาขาวิชา...วิทยาศาสตร์คอมพิวเตอร์... ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก...
ปีการศึกษา.....2553..... ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

5070457021 : MAJOR COMPUTER SCIENCE

KEYWORDS : LINK FARM / WEB SPAM / INFERENCE GRAPH GRAMMAR / WEB GRAPH / WEB SPAM DETECTION

WUTTICHAI WONGSARASIN : LINK FARM DETECTION USING GRAPH GRAMMAR INFERENCE. THESIS ADVISOR : ASST. PROF. ATHASIT SURARERKS, Ph.D., THESIS CO-ADVISOR : ASST. PROF. ARNON RUNGSAWANG, Ph.D., 53 pp.

Link farm is a group of web pages created for the purpose of increasing the rank scores. Accordingly, the high-rank-score pages would appear in the top rank of the search engine results. Thus, many researchers are focusing on improving methods to detect the link farm (also called web spam). In this thesis, the link farm graph grammar inference approach is introduced to recognize the link farm structures from web graph datasets. The graph grammar inference considers the link pattern specified by the labeled algorithm around the target host to distinguish the spam hosts from normal hosts. The output of the proposed algorithm is a grammar which represents a specific form of link farm. From the experimental results comparing with related approaches, the grammar obtained from the proposed inference approach can efficiently recognize link farms with high precision especially on the high rank score spam hosts environment.

Department : Computer Engineering	Student's Signature
Field of Study : Computer Science	Advisor's Signature
Academic Year : 2010	Co-Advisor's Signature

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์ และความช่วยเหลืออย่างดียิ่งจาก ผู้ช่วยศาสตราจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์ และผู้ช่วยศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง อาจารย์ที่ปรึกษา ที่ให้ข้อคิด ข้อเสนอแนะ คำปรึกษา ตลอดจนตรวจทานแก้ไขส่วนที่บกพร่องต่างๆ ในวิทยานิพนธ์ฉบับนี้ ขอขอบพระคุณอาจารย์ที่ปรึกษาทั้งสองท่านเป็นอย่างสูงที่ให้ความช่วยเหลือ เมตตา และโอกาสดีๆ แก่ผู้วิจัยเสมอมา

ขอขอบพระคุณ ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ประธานกรรมการสอบวิทยานิพนธ์ และรองศาสตราจารย์ ดร.เยาวดี เต็มธนาภัทร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ กรรมการสอบวิทยานิพนธ์ ที่กรุณาให้คำแนะนำ แก้ไขวิทยานิพนธ์ให้มีคุณภาพยิ่งขึ้น รวมทั้งขอขอบพระคุณคณาจารย์ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ประสิทธิประสาทวิชาความรู้อันมีค่ายิ่งแก่ผู้วิจัย

สุดท้ายนี้ขอขอบพระคุณ บิดา มารดา ที่เป็นกำลังใจอันสำคัญยิ่ง และขอบคุณเพื่อนๆ พี่ๆ น้องๆ ทุกคน โดยเฉพาะสมาชิกห้องปฏิบัติการวิจัยทางวิศวกรรมระบบนับได้เชิงทฤษฎี (ELITE) ที่ได้คอยให้ความช่วยเหลือ ข้อเสนอแนะ และแก้ไขเอกสาร จนสามารถทำให้วิทยานิพนธ์เสร็จสมบูรณ์ได้ ซึ่งผู้วิจัยหวังเป็นอย่างยิ่งว่า งานวิจัยนี้จะเป็นประโยชน์ต่อผู้ที่สนใจ และหากมีข้อผิดพลาดประการใด ผู้วิจัยขออภัยมา ณ ที่นี้ด้วย

สารบัญ

บทที่ หน้า

บทที่ **หน้า**

สารบัญตาราง

สารบัญภาพ

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันโปรแกรมค้นหาบนอินเทอร์เน็ต (search engine) มีบทบาทสำคัญต่อผู้ใช้ที่ต้องการค้นหาข้อมูลบนอินเทอร์เน็ตซึ่งมีอยู่เป็นจำนวนมาก โปรแกรมค้นหาที่ได้รับความนิยมและมีผู้ใช้จำนวนมากเช่น กูเกิล (Google) ยาฮู (Yahoo) บิง (Bing) เป็นต้น โปรแกรมค้นหาเหล่านี้จะทำการค้นคืน (retrieve) ข้อมูลข่าวสารบนโครงข่ายอินเทอร์เน็ต (world wide web: www) ให้ตรงกับความต้องการของผู้ใช้ โดยหลักการทำงานของโปรแกรมค้นหานี้ จะทำการรับคำค้น (query) จากผู้ใช้ และโปรแกรมค้นหาจะใช้อัลกอริทึมสำหรับการค้นคืน โดยค้นคืนเฉพาะข้อมูล และเว็บเพจ (web page) ที่เกี่ยวข้องกับคำค้นซึ่งถูกส่งมาจากผู้ใช้ และแสดงผลลัพธ์การค้นคืนให้ตรงกับความต้องการของผู้ใช้ให้มากที่สุด ในการค้นคืนแต่ละครั้งนั้น จะได้ข้อมูลเว็บเพจเป็นจำนวนมาก หลักการของการจัดลำดับผลลัพธ์การค้นคืนของโปรแกรมค้นหาบนอินเทอร์เน็ตคือ เว็บเพจที่มีความเกี่ยวข้องกับคำค้นมากที่สุด จะปรากฏอยู่ก่อนเว็บเพจที่มีความเกี่ยวข้องน้อยกว่า แต่เนื่องจากบนโครงข่ายอินเทอร์เน็ตมีข้อมูลเว็บเพจอยู่เป็นจำนวนมาก โปรแกรมค้นหาตัวอย่างเช่น กูเกิล ได้ออกแบบโปรแกรมค้นหา โดยใช้ปัจจัยต่างๆ เข้าร่วมกับการจัดลำดับ เช่น ความนิยมของเว็บเพจ (จำนวนลิงก์จากเว็บเพจอื่นที่ชี้เข้ามายังเว็บเพจนั้นๆ) คุณภาพของเว็บเพจ เป็นต้น ซึ่งผลการค้นคืนในหน้าแรกจะเป็นข้อมูลที่ผู้ใช้มักสนใจและเลือกที่จะเข้าไปยังหน้าเว็บเพจมากด้วยเช่นกัน ด้วยเหตุนี้ ผู้สร้างเว็บไซต์ที่ต้องการให้โปรแกรมค้นหา ค้นเจอเว็บไซต์ของตนและจัดเรียงอยู่ในหน้าแรกหรืออันดับต้นๆ ของการค้นคืน จะต้องพยายามสร้างเว็บของตนให้มีลักษณะตรงตามคำแนะนำ (guideline) ของโปรแกรมค้นหานั้นๆ

เทคนิคที่ใช้ในการสร้างเว็บไซต์ให้มีลักษณะที่โปรแกรมค้นหาจัดอันดับไว้ในลำดับต้นๆ มีอยู่หลายวิธี โดยเรียกเทคนิคต่างๆ เหล่านี้ว่า การปรับปรุงหน้าเว็บให้เหมาะสมกับแนวทางการค้นคืนของโปรแกรมค้นหา (Search Engine Optimization: SEO) เป็นการจัดทำ ปรับปรุง ปรับแต่งหน้าเว็บไซต์ เพื่อให้การจัดอันดับการค้นคืนนั้นอยู่ในอันดับต้นๆ ของโปรแกรมค้นหา แต่อย่างไรก็ตามเทคนิคการปรับแต่งหน้าเว็บมีอยู่หลายวิธี ทั้งเป็นการปรับปรุงหน้าเว็บในทางในทางที่ถูกต้อง (white SEO) และในทางที่ไม่ถูกต้อง (black SEO) ซึ่งผู้สร้างเว็บไซต์ (web site) ในเชิงพาณิชย์หรือเว็บไซต์โฆษณา บางกลุ่มมักจะใช้การปรับปรุงหน้าเว็บในทางที่ไม่ถูกต้อง ทั้งนี้เพื่อหวังผลให้เว็บไซต์ของตนปรากฏในลำดับต้นๆ ของผลลัพธ์การค้นคืนที่สูงกว่าปกติเพื่อให้มีผู้เข้าชมเว็บไซต์มากขึ้น เว็บไซต์ที่ถูกสร้างขึ้นโดยวิธีที่ไม่ถูกต้องนี้เรียกว่า เว็บสแปม (web spam)

เทคนิคการทำเว็บสแปมสามารถแบ่งออกเป็น 2 ประเภท คือ การทำสแปมเนื้อหา (content spam) และการทำสแปมโครงสร้างลิงก์ (link spam) [1] การทำสแปมเนื้อหา คือการสร้างเว็บเพจโดยบิดเบือนเนื้อหาให้ตรงกับคำค้นมากที่สุด เช่น การซ้ำคำค้น (repetition) เพื่อให้เว็บเพจดังกล่าวมีค่าความเกี่ยวข้องกับคำค้นเพิ่มขึ้น การใส่พจนานุกรม (dictionary) เพื่อให้เนื้อหาตรงกับคำค้นที่หลากหลาย เป็นต้น ส่วนการทำสแปมโครงสร้างลิงก์ คือการสร้างกลุ่มของเว็บเพจ และโครงสร้างลิงก์ให้มีการเชื่อมโยงกันภายในกลุ่มอย่างหนาแน่นเพื่อให้โปรแกรมค้นหาเข้าใจว่าเป็นเว็บที่มีความนิยมเพราะมีลิงก์จากเว็บเพจอื่นชี้มายังเพจนั้น ทั้งนี้เพื่อประโยชน์ในการถูกจัดอันดับแบบอิงโครงสร้างลิงก์ ซึ่งจะเรียกกลุ่มของเว็บเพจ และโครงสร้างลิงก์ที่เชื่อมถึงกันอย่างหนาแน่นที่ถูกสร้างขึ้นมาอย่างจงใจนี้ว่า ลิงก์ฟาร์ม (link farm) การทำเว็บสแปมไม่ว่าจะด้วยวิธีใดก็ตาม จะทำให้คุณภาพ และความน่าเชื่อถือของโปรแกรมค้นหาลดลง นอกจากนี้เว็บเพจบนอินเทอร์เน็ตนั้นมีจำนวนมากมายมหาศาล ซึ่งยากต่อการจำแนกเว็บสแปม และเว็บที่ไม่ใช่สแปมหรือเว็บปกติ (normal web) ออกจากกัน ทำให้นักวิจัยระบบสืบค้นข้อมูลจำนวนมากหันมาสนใจงานวิจัยทางด้านการตรวจจับ และลดผลกระทบในการจัดอันดับผลลัพธ์ของการค้นคืนจากเว็บสแปม เพื่อให้ได้ผลการค้นคืนที่มีประสิทธิภาพ และมีค่าความแม่นยำสูง

งานวิจัยทางด้านการตรวจจับเว็บสแปมนี้ มีนักวิจัยจำนวนมากให้ความสนใจอย่างกว้างขวางทั้งการตรวจจับเชิงเนื้อหา (content-based) [2] การตรวจจับเชิงโครงสร้างลิงก์ (link-based) [3-4] และการตรวจจับโดยวิเคราะห์เชิงเนื้อหาและเชิงโครงสร้างลิงก์เข้าด้วยกัน [5] นอกจากนี้ยังมีงานวิจัยที่นำไวยากรณ์กราฟ (graph grammar) มาใช้อธิบายโครงสร้างลิงก์ [6] ซึ่งสามารถใช้อธิบายลักษณะโครงสร้างลิงก์ฟาร์มของเว็บสแปมได้ แต่ไวยากรณ์กราฟดังกล่าวยังสามารถใช้อธิบายเว็บปกติบางเว็บได้เช่นกัน ดังนั้นในงานวิจัยนี้จึงได้สนใจการสร้างไวยากรณ์กราฟลิงก์ฟาร์มที่มีประสิทธิภาพเพื่ออธิบายโครงสร้างลิงก์ฟาร์มของเว็บสแปมได้อย่างถูกต้อง โดยใช้กระบวนการอนุมานไวยากรณ์กราฟ (graph grammar inference) ซึ่งการใช้การอนุมานไวยากรณ์กราฟเพื่ออธิบายโครงสร้างเว็บสแปม ไม่เพียงสามารถตรวจจับเว็บสแปมได้เท่านั้น แต่ยังได้ไวยากรณ์กราฟที่แสดงถึงลักษณะเฉพาะ (characteristic) ของโครงสร้างลิงก์ฟาร์มอีกด้วย นอกจากนี้โครงสร้างของลิงก์จะเป็นตัวบ่งบอกถึงความผิดปกติในการสร้างเว็บเพจทั่วไป เพราะลิงก์ฟาร์มส่วนใหญ่จะถูกสร้างขึ้นอย่างจงใจ ทำให้มีรูปแบบเฉพาะของโครงสร้างที่แตกต่างไปจากโครงสร้างของเว็บปกติทั่วไป ดังนั้นการใช้ไวยากรณ์กราฟลิงก์ฟาร์มที่ได้จากการอนุมาน จึงเป็นอีกวิธีการหนึ่งในการตรวจจับเว็บสแปมที่มีประสิทธิภาพได้

ส่วนงานวิจัยทางด้านการอนุมานไวยากรณ์กราฟนี้ มีนักวิจัยบางกลุ่มที่สนใจงานวิจัยทางด้านนี้เรื่อยมา เริ่มจากการศึกษาการอนุมานไวยากรณ์กราฟด้วยวิธีแทนที่ด้วยเส้นเชื่อม

(hyper-edge replacement graph grammar) [7] ทดลองบนกราฟที่ไม่มีทิศทาง (undirected graph) และไม่มีป้ายกำกับโหนด (unlabeled node) ซึ่งสามารถใช้หาไวยากรณ์กราฟที่อธิบายกราฟสองส่วนสมบูรณ์ (complete bipartite graph) ได้ จากนั้นได้มีการเสนอปัญหาการอนุมานโดยใช้ทฤษฎีความน่าจะเป็นในการอนุมานด้วยวิธีแทนที่ด้วยเส้นเชื่อมบนไวยากรณ์กราฟไม่พึ่งบริบท (context free graph grammar) [8] และพบว่าสามารถหาความน่าจะเป็นของกฎที่ใช้ในการอธิบายกราฟได้ นอกจากนี้ยังมีการนำเสนออัลกอริทึมการอนุมานไวยากรณ์กราฟด้วยวิธีแทนที่ด้วยโหนด (node replacement graph grammar) [9] โดยหลักการทำงานของอัลกอริทึมเริ่มจากการหากราฟย่อย (sub graph) ที่เกิดขึ้นซ้ำๆ กันบนกราฟตั้งต้น (input graph) และแทนที่กราฟที่ซ้ำกันเหล่านั้นด้วยโหนดเพียงหนึ่งโหนด และทำการทดลองการอนุมานกับโครงสร้างทางเคมี และจากนั้นงานวิจัยนี้ได้ถูกพัฒนาต่อมาโดยลดข้อจำกัดบางประการเพื่อใช้กับโครงสร้างกราฟที่ซับซ้อนยิ่งขึ้น โดยสามารถอนุมานไวยากรณ์กราฟจากโครงสร้างข้อมูลที่เป็นภาษาเอ็กซ์เอ็มแอล (Extensible Markup Language: XML) [10] แต่อย่างไรก็ตาม อัลกอริทึมการอนุมานไวยากรณ์กราฟด้วยการแทนที่โหนดและเส้นเชื่อมดังกล่าว ยังมีข้อจำกัดหากนำมาใช้อนุมานโครงสร้างลิงก์ฟาร์มบนเว็บกราฟ (web graph) ทั้งนี้เพราะเว็บกราฟมีลักษณะเป็นโครงสร้างที่ซับซ้อน และมีขนาดใหญ่มาก อีกทั้งข้อจำกัดในเรื่องของป้ายกำกับเส้นเชื่อม (edge label) และป้ายกำกับโหนด (node label) ซึ่งบนเว็บกราฟยังไม่มีป้ายกำกับเหล่านี้ ดังนั้นในงานวิจัยนี้จึงนำเสนอการอนุมานไวยากรณ์กราฟ และนำไวยากรณ์กราฟที่ได้มาตรวจจับลิงก์ฟาร์มต่อไป

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อศึกษา และออกแบบอัลกอริทึมการอนุมานไวยากรณ์กราฟสำหรับตรวจจับเว็บสแปมในลักษณะโครงสร้างที่เป็นลิงก์ฟาร์ม

1.3 ขอบเขตของการวิจัย

1) ศึกษาแนวทางการกำหนดป้ายกำกับโหนดและเส้นเชื่อมโดยใช้ฟังก์ชันลักษณะเฉพาะ
2) เสนออัลกอริทึมการอนุมานไวยากรณ์กราฟสำหรับลิงก์ฟาร์ม
3) เสนอไวยากรณ์กราฟสำหรับลิงก์ฟาร์มจากการอนุมานไวยากรณ์กราฟ
4) ทดสอบไวยากรณ์กราฟจากการอนุมานเพื่อตรวจจับลิงก์ฟาร์มอื่นและเปรียบเทียบประสิทธิภาพการตรวจจับเว็บสแปมกับงานวิจัยอื่นที่เกี่ยวข้อง
5) ชุดข้อมูลที่ใช้ในการทดลองเป็นชุดข้อมูลโฮสต์กราฟ (host graph) จากโดเมนประเทศอังกฤษ (.uk) ซึ่งรวบรวมโดยฝ่ายวิจัยยาฮู [11]

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) ได้อัลกอริทึมสำหรับการอนุมานไวยากรณ์กราฟบนเว็บกราฟ
2) ได้ไวยากรณ์กราฟที่แสดงถึงลักษณะเฉพาะทางโครงสร้างของลิงก์ฟาร์ม
3) สามารถตรวจจับลิงก์ฟาร์มโดยวิธีการอนุมานไวยากรณ์กราฟได้

1.5 วิธีดำเนินการวิจัย

1) ศึกษางานวิจัยและข้อมูลเอกสารที่เกี่ยวข้องกับลิงก์ฟาร์มและเว็บสแปม การตรวจจับลิงก์ฟาร์มและเว็บสแปมประเภทต่างๆ และการอนุมานไวยากรณ์กราฟ
2) วิเคราะห์หลักการและเทคนิคการตรวจจับลิงก์ฟาร์มและเว็บสแปม รวมทั้งวิเคราะห์อัลกอริทึมการอนุมานไวยากรณ์กราฟ
3) ออกแบบอัลกอริทึมการอนุมานไวยากรณ์กราฟสำหรับตรวจจับลิงก์ฟาร์ม
4) ทดสอบประสิทธิภาพอัลกอริทึมการอนุมานไวยากรณ์กราฟสำหรับตรวจจับลิงก์ฟาร์ม
5) ปรับปรุงไวยากรณ์กราฟสำหรับตรวจจับลิงก์ฟาร์ม
6) สรุปผลการวิจัยและตีพิมพ์ผลงานวิจัย
7) เรียบเรียงและจัดทำวิทยานิพนธ์

1.6 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์

1) "Characteristics of Link Farm for Graph Grammar Inference" โดย วุฒิชัย วงศ์สารสิน อานนท์ รุ่งสว่าง และอรรถสิทธิ์ สุรฤกษ์ ในการประชุมวิชาการ The 6th International Joint Conference on Computer Science and Software Engineering (JCSSE2009) ณ โรงแรมลากูนาบีชรีสอร์ท ภูเก็ต ระหว่างวันที่ 13-15 พฤษภาคม พ.ศ.2552
2) "Link Farm Formalization by Graph Grammar ". โดย วุฒิชัย วงศ์สารสิน อานนท์ รุ่งสว่าง และอรรถสิทธิ์ สุรฤกษ์ ในการประชุมวิชาการ The Conference on Knowledge and Smart Technologies (KST-2009) ณ มหาวิทยาลัยบูรพา ชลบุรี ระหว่างวันที่ 24-25 กรกฎาคม พ.ศ.2552

3) "Web Spam Recognition by Edge Label" โดย วุฒิชัย วงศ์สารสิน อานนท์ รุ่งสว่าง และอรรถสิทธิ์ สุรฤกษ์ ในการประชุมวิชาการ The 14th International Annual Symposium on Computational Science and Engineering (ANSCSE14) ณ มหาวิทยาลัยแม่ฟ้าหลวง เชียงราย ระหว่างวันที่ 23-26 มีนาคม พ.ศ.2553

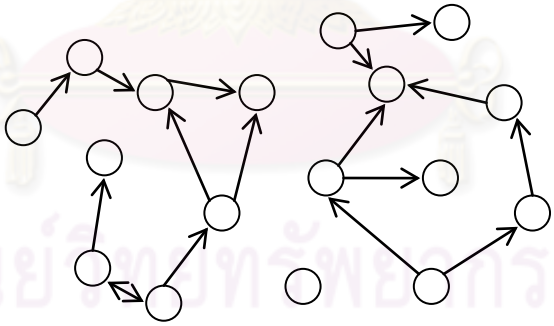
บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 เว็บกราฟ

เว็บกราฟ ถูกใช้เป็นแบบจำลองเครือข่ายของเว็บเพจ หรือโครงข่ายบนอินเทอร์เน็ต ประกอบด้วย โหนด (node) ที่แสดงถึงหน้าเว็บเพจที่มีการเชื่อมโยงถึงกันโดยเส้นเชื่อม (edge) หรือลิงก์ (link) เชื่อมถึงกันอย่างซับซ้อน มีงานวิจัยจำนวนหนึ่ง [3,6,12] ได้นิยามแบบจำลองเว็บกราฟขึ้นเพื่อความสะดวกในการใช้อธิบายความสัมพันธ์ต่างๆ ระหว่างเว็บเพจใดๆ บนอินเทอร์เน็ต ดังนิยามที่ 2.1

นิยามที่ 2.1 เว็บกราฟ คือ *กราฟระบุทิศทาง (directed graph) โดย G = (V, E)*
เมื่อ *V แทนเซตของโหนดที่แสดงถึงเว็บเพจ*
E แทนเซตของเส้นเชื่อมระหว่างโหนด
เส้นเชื่อม (a, b) ใน E หมายถึง เว็บเพจ a ชี้ไปยังเว็บเพจ b ทั้งนี้ต้องไม่เป็นเส้นวงวน (self loop) ที่ชี้เข้าหาโหนดตัวเอง



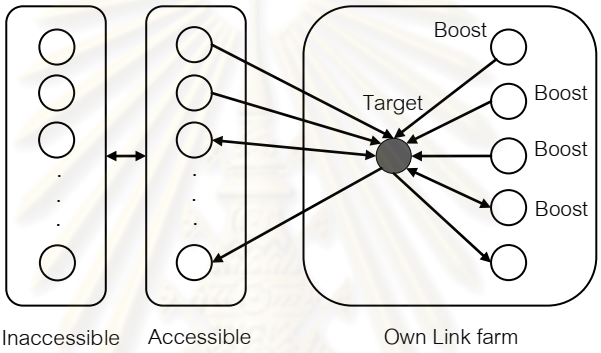
รูปที่ 2.1 ตัวอย่างแบบจำลองเว็บกราฟ

ตัวอย่างแบบจำลองเว็บกราฟสามารถแสดงได้ดังรูปที่ 2.1 โดยแต่ละโหนดอาจประกอบไปด้วยเส้นเชื่อมที่ชี้เข้าหาเรียกว่า อินลิงก์ (in-link) ส่วนเส้นเชื่อมที่ชี้ออกเรียกว่า เอาท์ลิงก์ (out-link) ในกรณีที่โหนดสองโหนดใดๆ มีลิงก์ที่ชี้เข้าหากันและกัน เส้นเชื่อมนั้นจะเรียกว่า ลิงก์สองทาง (reciprocity-link) จะเรียกจำนวนอินลิงก์ของโหนดใดๆ ว่า อินดีกรี (in-degree) เรียกจำนวนของ

เอาท์ลิงก์ว่า เอาท์ดีกรี (out-degree) และโหนดที่มีเส้นเชื่อมประชิดกันจะเรียกว่า โหนดเพื่อนบ้าน (neighbor node) ของกันและกัน

2.1.2 ลิงก์ฟาร์ม

ลิงก์ฟาร์ม คือกลุ่มของเว็บเพจที่มีจำนวนเส้นเชื่อมถึงกันอย่างหนาแน่น ลิงก์ฟาร์มอาจจะเกิดขึ้นได้ตามธรรมชาติจากกลุ่มของสังคมเว็บ (web communities) หรือเกิดขึ้นโดยความจงใจของผู้สร้างด้วยเทคนิคการทำโครงสร้างสแปม การสร้างลิงก์ฟาร์มแบบหลังนี้มีจุดประสงค์เพื่อเพิ่มความสำคัญให้กับเว็บเพจเป้าหมาย (target web page) ให้มีคะแนนการจัดอันดับอิงโครงสร้างลิงก์ที่มากกว่าปกติ ทำให้ผลการค้นคืนอยู่ในอันดับที่สูงขึ้นกว่าที่ควรจะเป็นในโปรแกรมค้นหา โดยสามารถแสดงแบบจำลองเว็บกราฟสำหรับการสร้างลิงก์ฟาร์ม [1] ได้ดังรูปที่ 2.2



รูปที่ 2.2 แบบจำลองเว็บกราฟสำหรับการสร้างลิงก์ฟาร์ม

จากรูปที่ 2.2 สามารถแบ่งประเภทของเว็บเพจตามมุมมองของผู้สร้างสแปม ได้เป็น 3 กลุ่มดังนี้

1. *กลุ่มเว็บเพจที่เป็นของผู้สร้างสแปม* (own link farm) คือกลุ่มเว็บเพจที่อยู่ภายใต้การควบคุมของผู้สร้างสแปม ซึ่งสามารถสร้างลิงก์ให้ชี้เข้าหาเว็บเพจเป้าหมายที่ต้องการได้ ภายในลิงก์ฟาร์มประกอบไปด้วยเว็บเพจหรือโฮสต์ที่เป็นเป้าหมายในการเพิ่มคะแนนการจัดอันดับ และกลุ่มของเว็บเพจ หรือโฮสต์ที่เพิ่มคะแนนให้กับเว็บเพจเป้าหมาย จะเรียกว่า บูซท์เพจ (boost pages) หรือ ซัพพอร์ตเตอร์ (supporters)
2. *กลุ่มเว็บเพจที่สามารถเข้าถึงได้* (accessible) คือกลุ่มของเว็บเพจที่ผู้สร้างสแปมสามารถเข้าไปแก้ไขหรือเพิ่มเติมเนื้อหาของเว็บเพจนั้นให้มีลิงก์ชี้เข้ามายังเว็บเพจเป้าหมายของตนได้ ยกตัวอย่างเช่น กลุ่มเว็บเพจประเภท บล็อก (blog) เว็บบอร์ด (web board) เป็นต้น

3. *กลุ่มเว็บเพจที่เข้าถึงไม่ได้* (inaccessible) เป็นกลุ่มเว็บเพจที่ผู้สร้างสแปมไม่สามารถเข้าไปแก้ไข หรือสร้างลิงก์เพื่อชี้มายังเว็บเพจของตนได้

2.1.3 **เพจแรงค์** (PageRank)

เพจแรงค์ [13] คืออัลกอริทึมการจัดอันดับความสำคัญให้กับเว็บเพจทั้งหมดบนเว็บกราฟ โดยมีหลักการคำนวณการถ่ายเทคะแนนตามความนิยมของเว็บ กล่าวคือ เว็บเพจใดมีอินลิงก์ที่ชี้เข้ามามากก็จะมีคะแนนความนิยมสูง ทำให้มีโอกาสอยู่ในอันดับต้นๆ ของผลการค้นคืน ปัจจุบันเพจแรงค์เป็นเครื่องมือที่นิยมใช้ในการจัดอันดับผลการค้นคืนมาก โดยได้ถูกนำมาใช้ในโปรแกรมค้นหาที่มีชื่อเสียงและนิยมมาก คือ กูเกิล

หลักการคำนวณคะแนนเพจแรงค์ของเว็บเพจใดๆ จะใช้ข้อมูลอินดีกรี เอาท์ดีกรี และใช้แบบจำลองการเดินสุ่ม (random walk) โดยใช้ความน่าจะเป็นในการกระโดดจากเว็บเพจหนึ่งสู่อีกเว็บเพจหนึ่งเท่าๆ กันทั้งเว็บกราฟ สามารถแสดงเป็นสมการการคำนวณได้ตามนิยามที่ 2.2

นิยามที่ 2.2 *เพจแรงค์ คือคะแนนการจัดลำดับที่กำหนดให้กับเว็บเพจใดๆ บนเว็บกราฟ คำนวณจากคะแนนเพจแรงค์ของอินลิงก์หารด้วยเอาท์ดีกรีของเว็บเพจที่เป็นสมาชิกของอินลิงก์เว็บเพจนั้นบวกกับคะแนนการสุ่มกระโดดไปยังเว็บเพจอื่น โดย*

$$r(u) = \delta \sum_{v \in I_u} \frac{r(v)}{O_v} + \frac{(1-\delta)}{n}$$

เมื่อ $r(u)$ *คือค่าคะแนนเพจแรงค์ของเว็บเพจ* u

$r(v)$ *คือค่าคะแนนเพจแรงค์ของเว็บเพจ* v

O_v *คือ จำนวนของเอาท์ดีกรีของเว็บเพจ* v

n *คือจำนวนเว็บเพจทั้งหมดบนเว็บกราฟ*

I_u *คือ เซตของเว็บเพจที่มีอินลิงก์ชี้ไปยังเว็บเพจ* u

δ *คือค่าคงที่ (damping factor) นิยมใช้* $\delta = 0.85$

2.1.4 **ทรังค์เคทเพจแรงค์** (Truncated PageRank)

ทรังค์เคทเพจแรงค์ [14] คือฟังก์ชันการจัดอันดับความสำคัญเชิงโครงสร้างลิงค์ที่มีหลักการคำนวณเช่นเดียวกับเพจแรงค์ แต่จะใช้เทคนิคลดความสำคัญของโครงสร้างโหนดเพื่อนบ้านที่ประชิดกับโหนดเป้าหมายลง เพื่อลดผลกระทบจากการถ่ายเทคะแนนโดยลิงก์ฟาร์ม ในงานวิจัยนี้ยังได้เสนอการคำนวณค่าคงที่ (δ) แบบใหม่ เพื่อใช้ประกอบในการคำนวณคะแนนการจัดทรังค์เคทเพจแรงค์ด้วย

2.1.5 ไวยากรณ์กราฟ

ไวยากรณ์กราฟ หมายถึง ตัวแบบ กฎ หรือ กติกา ที่ใช้อธิบายความซับซ้อนของภาษากราฟ (graph language) ซึ่งมีพื้นฐานมาจากทฤษฎีภาษารูปนัย (formal language) โดยสามารถนิยามได้ดังต่อไปนี้

นิยาม 2.3 *ไวยากรณ์กราฟ คือไวยากรณ์ที่อยู่ในรูป G = (S, P)*
เมื่อ S แทน กราฟเริ่มต้น
P แทน เซตของกฎโปรดักชัน (production rules) โดยโปรดักชันอยู่ในรูปของ $a \longrightarrow b$
โดย a *และ* b *เป็นกราฟ*

ไวยากรณ์กราฟ ถูกนำมาใช้ในการอธิบายโครงสร้างของกราฟที่มีลักษณะซ้ำๆ กัน ทั้งนี้เพื่อประโยชน์ในการลดขนาดของกราฟ และทำให้โครงสร้างของกราฟอยู่ในรูปแบบทั่วไปได้

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 อัลกอริทึมการตรวจจับเว็บสแปมเชิงโครงสร้างลิงก์

ในปี ค.ศ.2006 ลูคา เบคแชทติ และคณะ (Luca Becchetti et al.) [3] ได้นำเสนออัลกอริทึมการตรวจจับเว็บสแปมเชิงโครงสร้างลิงก์ โดยใช้การวิเคราะห์ทางสถิติ จากข้อมูลลักษณะเด่น (feature) ของเว็บเพจต่างๆ ได้แก่ ความสัมพันธ์ระหว่างดีกรี จำนวนเว็บเพจเพื่อนบ้าน คะแนนการจัดอันดับความสำคัญเชิงลิงก์ซึ่งประกอบไปด้วย เพจแรงค์ ทรัสต์แรงค์ และทรังค์เคทเพจแรงค์ โดยการทดลองได้นำลักษณะเด่นเหล่านี้มาวิเคราะห์ลักษณะเฉพาะของเว็บสแปม และนำมาตรวจจับเว็บสแปม โดยมีการวิเคราะห์แบบเดี่ยว และแบบรวมกลุ่มลักษณะเด่นต่างๆ เข้าด้วยกัน โดยรายงานว่าสามารถตรวจจับเว็บสแปมด้วยค่าความแม่นยำ (precision) เท่ากับ 80.4% และค่าความผิดพลาดแบบบวกลวง (false positive) เพียงแค่ 1.1%

2.2.2 การตรวจจับเว็บสแปมโดยเทคนิคเชิงโครงสร้างและเชิงเนื้อหา

ในปี ค.ศ.2007 คาร์ลอส คาสติโล และคณะ (Carlos Castillo et al.) [15] ได้เสนอระบบการตรวจจับเว็บสแปมโดยรวมเทคนิคการตรวจจับทั้งแบบโครงสร้างสแปม และแบบสแปมเนื้อหาเข้าด้วยกัน ซึ่งเทคนิคการตรวจจับสแปมโครงสร้างลิงก์นั้นได้ใช้ลักษณะเด่น เช่นเดียวกับงานวิจัยของ ลูคา เบคแชทติ และคณะ (Luca Becchetti et al.) [3] ที่กล่าวมาข้างต้น และในส่วนการตรวจจับสแปมเนื้อหา ได้ใช้ลักษณะเด่นจากงานวิจัยของ อเล็กซ์ซานดรอส นทูลลาส และคณะ

(Alexandros Ntoulas et al.) [2] เช่น จำนวนคำในเว็บเพจ จำนวนคำในหัวข้อ (title) และความยาวเฉลี่ยของคำ เป็นต้น จากนั้นใช้ตัวจำแนก (classifiers) แบบต้นไม้ตัดสินใจ (C4.5) เป็นตัวจำแนกเว็บสแปมออกจากเว็บปกติ โดยสามารถตรวจจับเว็บสแปมได้ด้วยค่าความแม่นยำถึง 88.4% แต่ก็ยังมีค่าความผิดพลาดแบบบวกลวงถึง 6.3%

2.2.3 การตรวจจับเว็บสแปมโดยไวยากรณ์กราฟ

ค.ศ.2008 เกียรติคุณ ชอบธรรม และคณะ (Kiattikun Chobtham et al.) [6] ได้เสนอแบบจำลองการคัดแยกลิงก์ฟาร์มในรูปแบบไวยากรณ์กราฟที่แปลงมาจากแบบจำลองลิงก์ฟาร์มที่เหมาะสม (optimal link farm model) จากนั้นใช้ไวยากรณ์กราฟที่สร้างตรวจจับลิงก์ฟาร์มด้วยการแจงส่วนไวยากรณ์กราฟ (graph grammar parser) ซึ่งผลการทดลองบนชุดข้อมูลทดสอบพบว่าไวยากรณ์กราฟที่สร้างขึ้นสามารถใช้อธิบายโครงสร้างเว็บสแปมได้ถึง 99% แต่ยังมีเว็บปกติบางเว็บที่ใช้ไวยากรณ์กราฟนี้อธิบายได้เช่นกัน นอกจากนี้ในงานวิจัย [16] ยังได้พัฒนาไวยากรณ์กราฟจากแบบจำลองลิงก์ฟาร์มที่เหมาะสมดังกล่าว พร้อมด้วยกฎตรรกศาสตร์ ในการตรวจจับลิงก์ฟาร์ม พบว่าสามารถตรวจจับลิงก์ฟาร์มในถังที่มีคะแนนเพจแรงค์สูง คือถังที่ 1 ถึง 5 จากทั้งหมด 10 ถังข้อมูล ดังแสดงในตารางที่ 2.1

ตารางที่ 2.1 ประสิทธิภาพการตรวจจับลิงก์ฟาร์มที่มีคะแนนเพจแรงค์ในถังที่ 1 ถึง 5

ถังที่	ค่าเรียกคืน	ค่าความแม่นยำ
1	0.7429	90.9090
2	1.9316	96.2963
3	5.4977	86.0465
4	4.8291	84.4155
5	6.9093	80.8695

จากตารางแสดงประสิทธิภาพการตรวจจับลิงก์ฟาร์มพบว่า ค่าความแม่นยำในการตรวจจับค่อนข้างสูงในทุกถัง แต่อย่างไรก็ตามค่าเรียกคืนในทุกถังอยู่ในระดับที่ต่ำ

2.2.4 การตรวจจับเว็บสแปมโดยวิธีแอนตี้ทรัสต์

ปี ค.ศ.2006 ไวเจย์ คริสนัน และคณะ (Vijay Krishnan et al.) [17] เสนอเครื่องมือการเลือกเซตสแปมเพจเริ่มต้น (seed) ซึ่งถูกตัดสินโดยผู้เชี่ยวชาญ โดยเลือกโครงสร้างลิงก์ของเว็บและติดป้ายฉลากให้เป็นเซตเริ่มต้นเพื่อตรวจจับสแปมเพจอื่น ผลการทดลองพบว่าสามารถตรวจจับลิงก์ฟาร์มได้ด้วยค่าความแม่นยำที่สูงกว่าวิธีทรัสแรงค์ [21] ขั้นตอนการทำงานสามารถแสดงได้ดังนี้

1) กำหนดเซตสแปมเริ่มต้นและติดฉลากด้วยผู้เชี่ยวชาญ โดยเลือกสแปมที่มีคะแนนเพจแรงค์สูง

2) คำนวณเมทริกซ์สลับเปลี่ยน (transpose) ของเมทริกซ์เว็บกราฟในรูปของเลขฐานสอง (binary webgraph)

3) คำนวณอัลกอริทึมคำนวณเพจแรงแบบเอนเอียง (biased PageRank) กับเมทริกซ์ T

4) เรียงลำดับคะแนนเพจแรงค์ตามลำดับมากไปน้อย ซึ่งจะเป็นการเรียงลำดับจากการการประมาณค่าความเป็นสแปม สามารถใช้ค่าขีดแบ่ง (threshold) ในการแบ่งแยกสแปมเพจออกจากเพจปกติทั่วไปได้

2.2.5 การตรวจจับเว็บสแปมโดยวิธีทรานส์ดักทิฟ

ในปี ค.ศ.2007 เดงยง ซัว และคณะ (Dengyong Zhou et al.) [18] ได้เสนอการตรวจจับลิงก์ฟาร์มโดยใช้วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) และใช้ชุดตัวอย่างทดสอบทั้งโฮสต์ สแปม (ตัวอย่างบวก) และโฮสต์ปกติ (ตัวอย่างลบ) เช่นเดียวกับงานวิจัยนี้ ผลการทดลองพบว่าสามารถตรวจจับเว็บสแปมได้ดีกว่าอัลกอริทึมแอนตี้ทรัสโดยเฉพาะในช่วงค่าการเรียกคืนสูงอัลกอริทึมการงานสามารถเขียนเป็นขั้นตอนได้ดังนี้

1) กำหนดให้เว็บกราฟ $G = (V, E)$ โดย V คือเซตของโหนดและ E คือเซตของเส้นเชื่อม โดยเว็บเพจ $S \subset V$ เป็นเว็บเพจที่ติดฉลากไว้แล้วว่าเป็นเว็บปกติหรือเว็บสแปม และสมมติให้กราฟมีการเชื่อมถึงกันอย่างหนาแน่น (strongly connected) แต่ถ้าไม่ใช่โครงสร้างดังกล่าว จะทำการย่อยโครงสร้างกราฟให้อยู่ในโครงสร้างที่เชื่อมโยงกันอย่างหนาแน่น สำหรับเว็บเพจที่ยังไม่ได้ติดฉลากไว้ในเซต V สามารถแสดงได้ดังขั้นตอนต่อไปนี้

2) กำหนดฟังก์ชันการเดินสุ่มโดยเลือกลิงก์ที่ชี้เข้าและสุ่มด้วยความน่าจะเป็น

$$p(u,v) = \frac{w(v,u)}{d^{-}(u)},$$

เมื่อ u และ v คือ โหนดใดๆ ใน V

$w(v,u)$ คือ น้ำหนักของเส้นเชื่อมระหว่าง v และ u

$d^{-}(u)$ คือ ผลรวมของน้ำหนักของเส้นเชื่อมที่มีโหนดปลายทางเป็น u

กำหนดให้ Π แทนเวกเตอร์ โดยที่

$$\sum_{u \in V} \Pi(u) p(u,v) = \Pi(v)$$

3) กำหนดให้ P เป็นเมทริกซ์ที่มีสมาชิกแต่ละตัวเป็น $p(u,v)$ และ Π เป็นเมทริกซ์ทแยงมุม (diagonal matrix) ที่มีสมาชิกเป็น $\Pi(u)$ ดังนั้นรูปแบบเมทริกซ์จะเป็น

$$L = \Pi - \omega \frac{\Pi P + P^T \Pi}{2}$$

เมื่อ ω คือค่าพารามิเตอร์ในช่วง]0,1[

4) กำหนฟังก์ชันก์ y บนโหนด V ด้วย y(v) = 1 หรือ -1 ถ้าเว็บเพจ v ถูกติดฉลากด้วยเว็บปกติหรือเว็บสแปมตามลำดับ และจะถูกติดฉลากด้วย 0 ถ้า v ไม่มีป้ายติดฉลาก (unlabeled) ดังการแก้สมการเชิงเส้น

$$L\varphi = \Pi y$$

จากนั้นจำแนกเว็บเพจ v แต่ละเว็บที่ยังไม่ได้ติดป้ายฉลากด้วยฟังก์ชัน $\varphi(v)$

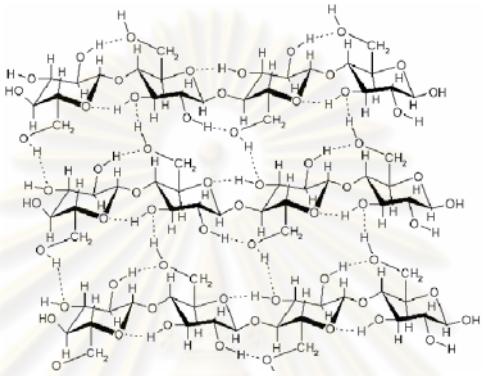
2.2.6 การอนุมานไวยากรณ์กราฟ

งานวิจัยทางด้านการอนุมานไวยากรณ์กราฟนั้น มีนักวิจัยจำนวนหนึ่งได้สนใจและนำเสนออัลกอริทึมสำหรับการอนุมานกราฟ เพื่อหาไวยากรณ์กราฟที่แสดงถึงลักษณะทางโครงสร้างของกราฟ ซึ่งสามารถนำไวยากรณ์กราฟดังกล่าว มาสร้างกราฟใหม่ที่มีลักษณะโครงสร้างที่เหมือนกับกราฟตั้งต้นได้ ทั้งนี้เพื่อประโยชน์ในการอธิบายสมบัติของกราฟในรูปแบบทั่วไปได้ นอกจากนี้ยังเป็นการลดขนาดของกราฟที่มีขนาดใหญ่ให้อยู่ในรูปไวยากรณ์ที่มีขนาดเล็กกว่า อีกทั้งยังแสดงถึงโครงสร้างของกราฟในรูปที่เข้าใจง่ายอีกด้วย

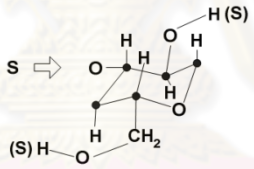
ในปี ค.ศ.2007 จาเคค คุกลุก (Jacek Kukluk) [19] ได้เสนออัลกอริทึมการอนุมานไวยากรณ์กราฟโดยการแทนที่ด้วยโหนดและเส้นเชื่อม โดยพัฒนามาจากอัลกอริทึมซับดิว (Subdue) ซึ่งเป็นอัลกอริทึมสำหรับหากราฟย่อย (sub graph) ที่ซ้ำกัน โดยอัลกอริทึมแบ่งเงื่อนไขการทำงานออกเป็นสองวิธี คือ ใช้การอนุมานโดยการแทนที่ด้วยโหนด หากตรวจพบการซ้อนทับ (overlap) กันหนึ่งโหนดของกราฟย่อย และจะใช้การอนุมานโดยการแทนที่ด้วยเส้นเชื่อม หาก

ตรวจพบการซ้อนทับกันสองโหนด (เส้นเชื่อม) ของกราฟย่อย ซึ่งการวัดประสิทธิภาพของอัลกอริทึมการอนุมานในงานวิจัยดังกล่าว จะใช้การสร้างกราฟจากไวยากรณ์เริ่มต้นที่ต้องการตรวจสอบ และนำกราฟนี้มาผ่านกระบวนการอนุมานด้วยอัลกอริทึมดังกล่าว และเปรียบเทียบไวยากรณ์เริ่มต้นกับไวยากรณ์ที่ได้จากการอนุมาน

งานวิจัยการอนุมานไวยากรณ์กราฟของจาเคค คุกลุกนี้ได้ทดลองการอนุมานกับโครงสร้างทางเคมี เช่นโครงสร้างทางเคมีของเซลลูโลสกับพันธะไฮโดรเจน ดังรูปที่ 2.3



(a) โครงสร้างทางเคมีของเซลลูโลสกับพันธะไฮโดรเจน



(b) ไวยากรณ์กราฟที่ได้จากอัลกอริทึมการอนุมานไวยากรณ์กราฟ

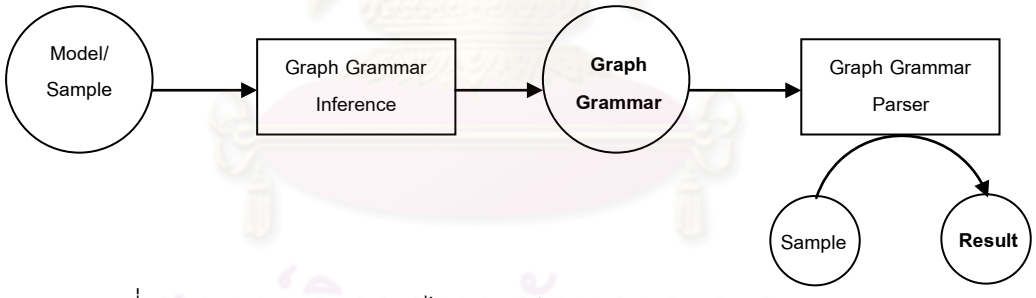
รูปที่ 2.3 ตัวอย่างโครงสร้างทางเคมีและไวยากรณ์กราฟ

จากรูปที่ 2.3 (a) จะเห็นได้ว่าโครงสร้างดังกล่าวมีรูปแบบของโครงสร้างย่อยที่เหมือนกันอยู่หลายตำแหน่ง ซึ่งสามารถใช้กระบวนการอนุมานเพื่อหาไวยากรณ์กราฟของโครงสร้างดังกล่าวได้ดังรูปที่ 2.3 (b) ไวยากรณ์กราฟดังกล่าว ประกอบด้วยกฎโปรดักชันเพียงหนึ่งกฎเท่านั้น ซึ่งสามารถใช้อธิบายโครงสร้างทั้งหมดของโครงสร้างทางเคมีระหว่างเซลลูโลสกับพันธะไฮโดรเจนได้ทั้งหมด

บทที่ 3
การอนุมานไวยากรณ์กราฟและการตรวจจับลิงก์ฟาร์ม

จากงานวิจัยที่เกี่ยวข้องพบว่าการตรวจจับเว็บสแปม โดยวิเคราะห์ลักษณะเด่นต่างๆ ของเว็บเพจนั้นสามารถจำแนกเว็บสแปมออกจากเว็บปกติได้ผลดีในระดับหนึ่ง แต่การตรวจสอบเช่นนี้ไม่ได้พิจารณาถึงลักษณะทางโครงสร้างที่แท้จริงของเว็บสแปม อีกทั้งการใช้ไวยากรณ์กราฟในการบรรยายลักษณะโครงสร้างเว็บสแปมที่เป็นลิงก์ฟาร์ม พบว่าสามารถใช้บรรยายลิงก์ฟาร์มได้ถึง 99% แต่อย่างไรก็ตาม มีเว็บปกติจำนวนหนึ่งที่ใช้ไวยากรณ์กราฟดังกล่าวอธิบายได้เช่นกัน การลดความผิดพลาดที่เกิดขึ้นจากการใช้ไวยากรณ์กราฟนั้น สามารถทำได้โดยสร้างไวยากรณ์ที่มีประสิทธิภาพ ดังนั้นในงานวิจัยนี้จึงมีแนวคิดในการนำเสนอการอนุมานไวยากรณ์กราฟที่แสดงถึงลักษณะโครงสร้างที่แท้จริงของเว็บสแปมที่อยู่รวมกันเป็นกลุ่มของลิงก์ฟาร์ม และสามารถใช้ไวยากรณ์ที่ได้จากการอนุมานนี้ในการตรวจจับเว็บสแปมหรือลิงก์ฟาร์มอื่นๆ ได้

กระบวนการอนุมานไวยากรณ์กราฟสำหรับตรวจจับลิงก์ฟาร์ม [20] เริ่มด้วยการอนุมานไวยากรณ์จากกราฟต้นแบบหรือกราฟตัวอย่างเริ่มต้น จะได้ไวยากรณ์กราฟที่แสดงถึงลักษณะโครงสร้างของกราฟต้นแบบ จากนั้นใช้ไวยากรณ์กราฟดังกล่าวตรวจจับลิงก์ฟาร์มด้วยวิธีการแจงส่วนไวยากรณ์กราฟ ดังแสดงกระบวนการในรูปที่ 3.1



รูปที่ 3.1 กระบวนการอนุมานไวยากรณ์กราฟ และการแจงส่วนจากตัวอย่าง

ในกระบวนการอนุมานไวยากรณ์กราฟบนเว็บกราฟนั้น จะแตกต่างจากการอนุมานไวยากรณ์กราฟอื่นที่มีโครงสร้างที่มีรูปแบบแน่นอน เช่น โครงสร้างทางเคมี เป็นต้น ทั้งนี้โครงสร้างที่ต้องการศึกษาในงานวิจัยนี้เป็นโครงสร้างลิงก์ฟาร์ม ที่มีความหลากหลายของโครงสร้าง เพราะเกิดจากการสร้างโดยผู้สร้างสแปมที่หลากหลาย อีกทั้งจำนวนเส้นเชื่อมของโครงสร้างลิงก์ฟาร์มมีจำนวนที่ไม่แน่นอน จึงต้องปรับอัลกอริทึมการอนุมานให้มีความยืนหยุ่นมากขึ้น และเนื่องจากอัลกอริทึมการอนุมานไวยากรณ์กราฟในงานวิจัยของจาเคค คุกลุก มีข้อจำกัดหลายประการหากนำมาใช้อนุมานบนเว็บกราฟ เพราะโหนดที่แสดงถึงเว็บเพจหรือโฮสต์นั้นมีลักษณะต่างกันทุกๆ เว็บเพจ อีกทั้งเส้นเชื่อมระหว่างโหนดไม่มีป้ายกำกับ ดังนั้นจึงมีแนวคิดในการอนุมานไวยากรณ์

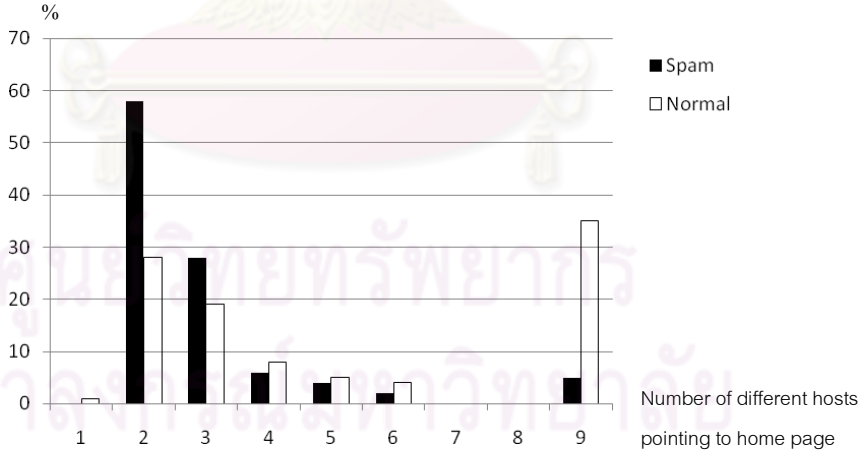
กราฟโดยกำหนดลักษณะเฉพาะของโหนด และเส้นเชื่อม เพื่อใช้เป็นป้ายกำกับในกระบวนการอนุมานไวยากรณ์กราฟ โดยลักษณะเฉพาะที่เพิ่มเติมเข้าไปบนเว็บกราฟนี้เป็นเซตของลักษณะเด่นที่บ่งบอกถึงลักษณะเฉพาะของโหนด และเส้นเชื่อมใดๆ ซึ่งการนำลักษณะเฉพาะกำหนดลงบนเว็บกราฟนี้เปรียบเสมือนการสร้างป้ายกำกับให้กับโหนด และเส้นเชื่อมใดๆ เพื่อนำมาประยุกต์ใช้ในกระบวนการอนุมานต่อไป

3.1 การวิเคราะห์ลักษณะเฉพาะของลิงก์ฟาร์ม

ลักษณะเฉพาะของลิงก์ฟาร์มโดยทั่วไปจะประกอบไปด้วยลักษณะเด่นหลายๆ ลักษณะรวมเข้าด้วยกัน งานวิจัยจำนวนมากมุ่งเน้นหาลักษณะเฉพาะของลิงก์ฟาร์มหรือเว็บสแปมที่แตกต่างออกจากเว็บปกติทั่วไป เพื่อสามารถใช้ลักษณะเฉพาะเหล่านี้ในการแบ่งหรือแยกแยะเว็บสแปมออกจากเว็บปกติ โดยการพิจารณาลักษณะเด่นต่างๆ สามารถแบ่งออกได้ใน 3 มุมมองดังนี้

3.1.1 ลักษณะเด่นเชิงดีกรี (degree-based)

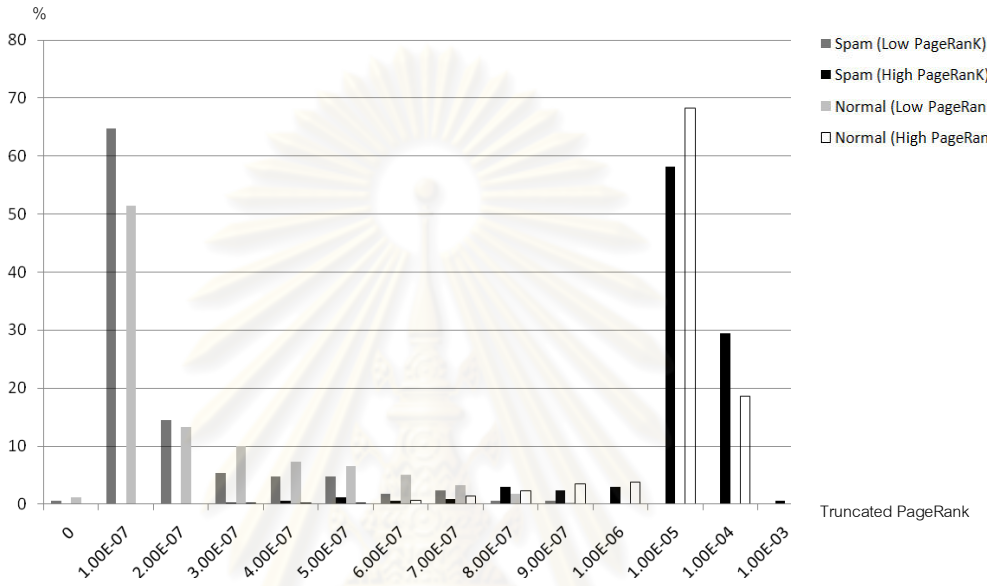
ลักษณะเด่นเชิงดีกรีนี้ จะวิเคราะห์ข้อมูลโครงสร้างลิงก์ โดยพิจารณาลักษณะต่างๆ ของอินลิงก์ เอาท์ลิงก์ และลิงก์สองทาง ซึ่งมีลักษณะเด่นอย่างหนึ่งที่น่าสนใจคือ จำนวนโฮสต์ที่แตกต่างกัน ซึ่งชี้เข้าหาเว็บเพจหลัก หรือโฮมเพจ (home page) ของโฮสต์สแปมเป้าหมายที่พิจารณา โดยส่วนใหญ่จะอยู่ในช่วง 0 ถึง 10 โฮสต์ ดังรูปที่ 3.2



รูปที่ 3.2 ฮิสโทแกรมแสดงจำนวนโฮสต์สแปมและโฮสต์ปกติในช่วงของจำนวนโฮสต์ที่แตกต่างกันซึ่งชี้เข้าหาโฮมเพจของโฮสต์สแปมเป้าหมาย

3.1.2 **ลักษณะเด่นเชิงการจัดอันดับผลการค้นคืน** (rank-based)

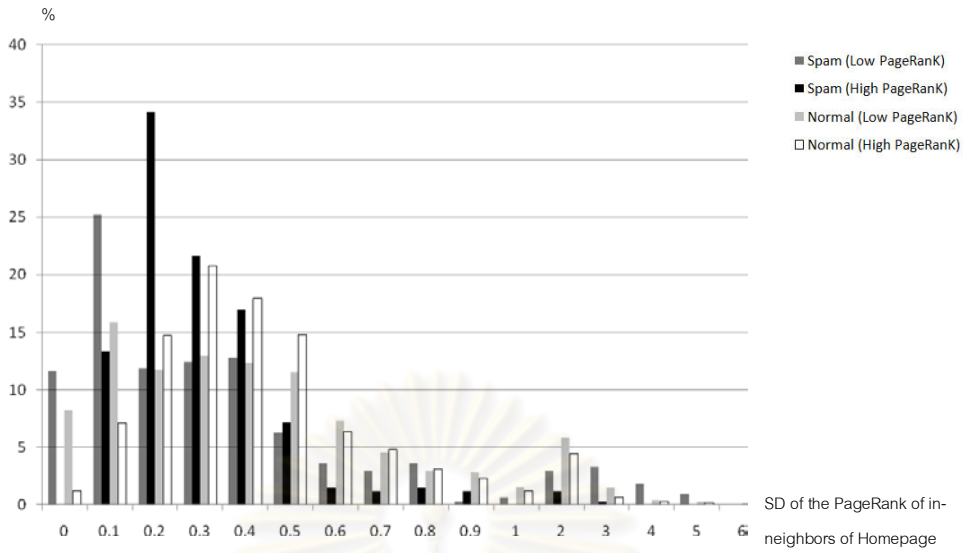
การพิจารณาลักษณะเด่นเชิงการจัดอันดับผลการค้นคืนนั้นจะพิจารณาอัลกอริทึมสำหรับการจัดอันดับที่นิยมใช้อยู่ในปัจจุบัน ได้แก่ เพจแรงค์ ทรังค์เคทเพจแรงค์ และทรัสต์แรงค์ [21] ซึ่งคะแนนทรังค์เคทเพจแรงค์เป็นลักษณะเด่นอย่างหนึ่งที่สอดคล้องกับคะแนนเพจแรงค์ คือ โฮสต์สแปมกลุ่มที่มีคะแนนเพจแรงค์สูงก็จะมีคะแนนทรังค์เคทเพจแรงค์ที่สูงด้วย ดังรูปที่ 3.3



รูปที่ 3.3 ฮิสโทแกรมแสดงจำนวนโฮสต์สแปม และโฮสต์ปกติ
ในช่วงคะแนนทรังเคทเพจแรงค์ต่างๆ

3.1.3 **ลักษณะเด่นเชิงโฮสต์เพื่อนบ้าน** (neighbor-based)

ลิงก์ฟาร์มถูกสร้างขึ้นอย่างจงใจ เพื่อให้มีโครงสร้างในการเพิ่มคะแนนการจัดอันดับให้กับเว็บเพจเป้าหมาย ดังนั้นลักษณะทางโครงสร้างการถ่ายเทลิงก์ของโฮสต์เพื่อนบ้านที่ชี้เข้าหาโฮสต์เป้าหมายนั้นมักจะมีโครงสร้างที่คล้ายกัน ซึ่ง แอนดราส เบนซ์ซูร์ และคณะ (Andras Benczur, et.al) [22] ได้ตั้งข้อสังเกตไว้ว่าเว็บเพจเพื่อนบ้านโดยรอบเว็บเพจเป้าหมายที่เป็นสแปมมักจะมีคะแนนเพจแรงค์ที่ใกล้เคียงกัน ทั้งนี้เพราะเว็บเพจเพื่อนบ้านเหล่านี้ถูกสร้างขึ้นอย่างจงใจเพื่อเพิ่มคะแนนเพจแรงค์ให้กับเว็บเพจเป้าหมาย ส่วนเว็บเพจเพื่อนบ้านโดยรอบเว็บเพจเป้าหมายที่เป็นเว็บปกติจะมีคะแนนเพจแรงค์ที่ค่อนข้างหลากหลาย ซึ่งเป็นลักษณะทั่วไปของการเชื่อมโยงกันระหว่างลิงก์บนอินเทอร์เน็ต ดังนั้นเมื่อพิจารณาคะแนนเพจแรงค์ของเว็บเพจเพื่อนบ้านโดยรอบโฮสต์เป้าหมายแล้วจะพบว่าค่าส่วนเบี่ยงเบนมาตรฐาน (standard deviation) ระหว่างคะแนนเพจแรงค์ของโฮสต์สแปมเป้าหมายกับคะแนนเพจแรงค์เฉลี่ยของโฮสต์เพื่อนบ้าน จะมีค่าอยู่ในช่วง 0 ถึง 0.4 แต่โฮสต์ปกติจะมีค่าเบี่ยงเบนมาตรฐานอยู่อย่างกระจายตัว ดังรูปที่ 3.4



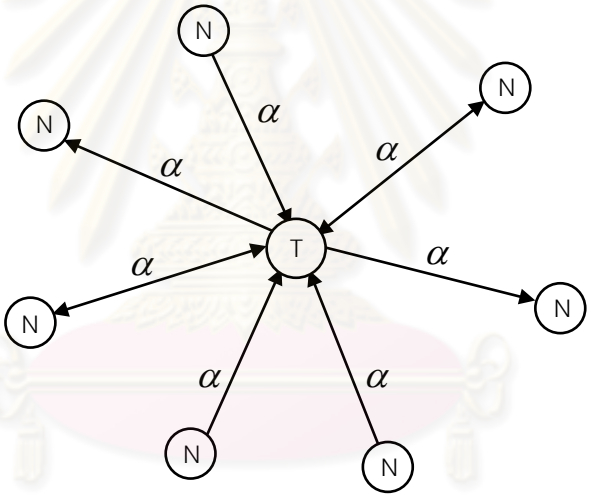
รูปที่ 3.4 ฮิสโทแกรมแสดงจำนวนโฮสต์สแปมและโฮสต์ปกติ
ในช่วงค่าเบี่ยงเบนมาตรฐานต่างๆ

นอกจากนี้หากพิจารณาค่าส่วนเบี่ยงเบนมาตรฐานที่แตกต่างกันระหว่างโฮสต์สแปมกลุ่มที่มีคะแนนเพจแรงค์สูงและกลุ่มที่มีคะแนนเพจแรงค์ต่ำพบว่ามีการกระจายตัวของค่าส่วนเบี่ยงเบนมาตรฐานที่แตกต่างกัน แสดงให้เห็นว่ากลุ่มโฮสต์ที่มีคะแนนเพจแรงค์สูงและคะแนนเพจแรงค์ต่ำมีโครงสร้างลิงก์ที่แตกต่างกัน ดังนั้นในการทดลองต่อไปจะแบ่งข้อมูลออกเป็น 10 ถังตามคะแนนเพจแรงค์ทั้งนี้ เพื่อประสิทธิภาพในการสร้างไวยากรณ์กราฟในช่วงคะแนนเพจแรงค์ที่ต่างกันออกไป

จากการวิเคราะห์ลักษณะเด่นของลิงก์ฟาร์มทั้ง 3 มุมมองแล้วพบว่าลักษณะเด่นต่างๆ สามารถแสดงความแตกต่างระหว่างโฮสต์สแปมกับโฮสต์ปกติได้ในบางช่วงลักษณะ งานวิจัยหลายงานวิจัยได้เลือกใช้ลักษณะเด่นจำนวนมากในการแบ่งแยกโฮสต์ทั้งสองออกจากกันซึ่งยิ่งถ้าใช้ลักษณะเด่นมากขึ้นการแบ่งแยกก็จะดีขึ้นตามด้วย แต่อย่างไรก็ตามการใช้ลักษณะเด่นจำนวนมากจะต้องยิ่งใช้เวลาในการวิเคราะห์หรือรวบรวมข้อมูลมากขึ้นด้วย เนื่องจากในงานวิจัยนี้เป็นการศึกษาโครงสร้างลิงก์ฟาร์ม ซึ่งเป็นโครงสร้างที่สร้างขึ้นเพื่อเพิ่มคะแนนการจัดอันดับ โดยถ่ายเทคะแนนมาจากเว็บเพจที่สร้างขึ้นในลิงก์ฟาร์ม ดังนั้นในงานวิจัยนี้จะเลือกใช้ลักษณะเด่นเพียง 2 ลักษณะซึ่งบ่งบอกถึงการถ่ายเทคะแนนการจัดอันดับมาร่วมในการสร้างไวยากรณ์กราฟด้วยนั่นคือ คะแนนเพจแรงค์ และคะแนนทรังเคทเพจแรงค์

3.2 การกำหนดป้ายกำกับเส้นเชื่อม

ในการทดลองใช้การอนุมานไวยากรณ์กราฟในการสร้างไวยากรณ์กราฟสำหรับตรวจจับโครงสร้างลิงก์ฟาร์ม ซึ่งการอนุมานไวยากรณ์กราฟนั้นมีหลักการคือ การสร้างไวยากรณ์กราฟจากโครงสร้างย่อย (sub structure) ที่มีโครงสร้างซ้ำๆ กัน แต่การจะตัดสินว่าโครงสร้างนั้นมีลักษณะเดียวกันหรือไม่นั้นอาจดูได้จากโครงสร้างของเส้นเชื่อมที่มีลักษณะการชี้เข้าหรือชี้ออกไปในทิศทางเดียวกัน แต่ในการตรวจจับลิงก์ฟาร์มนี้ การพิจารณาลักษณะการชี้เข้าหรือชี้ออกของลิงก์ยังไม่เพียงพอต่อการแยกแยะระหว่างโฮสต์สแปมและโฮสต์ปกติ ดังนั้นจึงต้องมีการกำหนดลักษณะเฉพาะลงบนโฮสต์กราฟ เพื่อสามารถแยกโฮสต์สแปมกับโฮสต์ปกติออกจากกันได้อย่างมีประสิทธิภาพ การกำหนดลักษณะเฉพาะในงานวิจัยนี้จะกำหนดลงบนเส้นเชื่อมของโฮสต์กราฟ โดยจะเรียกลักษณะเฉพาะที่ถูกกำหนดลงบนเส้นเชื่อมนี้ว่าป้ายกำกับเส้นเชื่อม ดังรูปที่ 3.5



เมื่อ α แทนป้ายกำกับของเส้นเชื่อม

T แทนโหนดเป้าหมายที่กำลังพิจารณา

N แทนโหนดเพื่อนบ้านของโหนด T

รูปที่ 3.5 การกำหนดป้ายกำกับเส้นเชื่อม

ป้ายกำกับเส้นเชื่อม (α) ที่กำหนดลงบนเส้นเชื่อมของเว็บกราฟแต่ละเส้น จะอยู่ในรูปเวกเตอร์ของเลขฐานสอง ซึ่งแต่ละจุดของเวกเตอร์จะเกิดจากการเปรียบเทียบค่าของลักษณะเด่นของโหนดต้นทางและโหนดปลายซึ่งมีเส้นเชื่อมเชื่อมโหนดทั้งสองอยู่ ดังนี้

กำหนดให้ป้ายกำกับเส้นเชื่อมประกอบไปด้วย *n* ค่า สามารถเขียนป้ายกำกับเส้นเชื่อมให้อยู่ในรูปของเวกเตอร์ได้คือ $\alpha = <\alpha_1, \alpha_2, \alpha_3, ..., \alpha_n>$ ซึ่งแต่ละค่าของเวกเตอร์ถูกกำหนดด้วยค่า 0 และ 1 ดังนี้

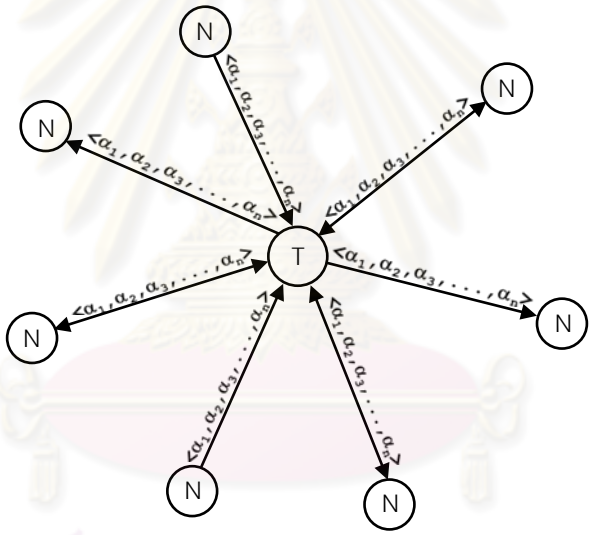
$$\alpha_i = \begin{cases} 0 & ; \text{ถ้า} \quad f_i(\mathrm{T}) \leq f_i(\mathrm{N}) \\ 1 & ; \text{ถ้า} \quad f_i(\mathrm{T}) > f_i(\mathrm{N}) \end{cases}$$

เมื่อ α_i คือ ลักษณะเด่นที่ *i*

f_i(T) คือ ค่าของลักษณะเด่นที่ *i* ของโฮสต์เป้าหมาย

f_i(N) คือ ค่าของลักษณะเด่นที่ *i* ของโฮสต์เพื่อนบ้าน

สามารถแสดงรูปการกำหนดป้ายกำกับเส้นเชื่อมได้ดังรูปที่ 3.6



รูปที่ 3.6 การกำหนดป้ายกำกับเส้นเชื่อมโดยลักษณะเด่น

ในส่วนต่อไปจะแสดงตัวอย่างการกำหนดลักษณะเด่นให้กับเส้นเชื่อมในระดับแรกของโฮสต์กราฟ

ตัวอย่างที่ 3.1 การกำหนดลักษณะเด่นให้กับเส้นเชื่อม

การกำหนดลักษณะเด่นจะใช้ชุดข้อมูลจากฝ่ายวิจัยยาฮู (Yahoo! Research) [11] ซึ่งเป็นโฮสต์กราฟที่ถูกติดฉลากประเภทโฮสต์ไว้แล้วโดยทีมนักวิจัย และอาสาสมัคร ในตัวอย่างนี้ได้เลือกโฮสต์เป้าหมายที่เป็นโฮสต์สแปมในลำดับที่ 5022 มีที่อยู่เว็บ (url) คือ www.car--loan.co.uk และใช้ลักษณะเด่นจำนวน 10 ลักษณะในการกำหนดลักษณะเฉพาะ ดังนี้

$$\alpha = <\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9, \alpha_{10}>$$

โดยลักษณะเด่นประกอบด้วยเวกเตอร์ของ 10 ลักษณะดังนี้

f_1: อินดีกรีของโฮมเพจ

f_2: เอาท์ดีกรีของโฮมเพจ

f_3: ค่าเฉลี่ยอินดีกรีของโฮมเพจเพื่อนบ้าน

f_4: ค่าเฉลี่ยเอาท์ดีกรีของโฮมเพจเพื่อนบ้าน

f_5: อัตราส่วนอินดีกรีต่อเอาท์ดีกรีของโฮมเพจ

f_6: อัตราส่วนดีกรีต่อดีกรีเฉลี่ยของโฮสต์เพื่อนบ้าน

f_7: จำนวนโฮสต์เพื่อนบ้านที่ชี้เข้าหาโฮสต์เป้าหมาย

f_8: จำนวนโฮสต์เพื่อนบ้าน ระยะทาง 2 จากโฮมเพจ

f_9: คะแนนทรังเคทเพจแรงค์ ระยะทาง 1 จากโฮมเพจ

f_{10}: คะแนนเพจแรงค์ของโฮมเพจ

จากนั้นพิจารณาการกำหนดลักษณะเฉพาะลงบนเส้นเชื่อมระหว่างโฮสต์เป้าหมายกับโฮสต์เพื่อนบ้าน ซึ่งพบว่าโหนดเป้าหมายดังกล่าวมีโฮสต์เพื่อนบ้าน 50 โฮสต์ สามารถกำหนดลักษณะเฉพาะลงบนเส้นเชื่อมทั้ง 50 เส้นได้ดังแสดงในตารางที่ 3.1

ตารางที่ 3.1 ลักษณะเฉพาะบนเส้นเชื่อมระหว่างโฮสต์เป้าหมายกับโฮสต์เพื่อนบ้าน

โฮสต์สแปมเป้าหมาย: www.car--loan.co.uk

ID	URL	Type	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α
252	all-car-insurances.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
253	all-carinsurance.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
254	all-carsinsurance.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
255	all-carsinsurances.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
260	allcar-insurance.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
261	allcar-insurances.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
263	allcars-insurances.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
264	allcarsinsurance.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
265	allcarsinsurances.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
2350	get4me.co.uk	อินลิงก์	1	0	0	0	1	1	1	1	1	1	B
2734	loan-searcher.co.uk	อินลิงก์	1	0	0	1	0	1	1	1	1	1	C
2893	mortgagejunction.co.uk	อินลิงก์	1	0	0	1	0	1	1	1	1	1	C
3787	www.0845assurance.co.uk	อินลิงก์	0	0	0	1	1	1	1	0	1	1	D
3788	www.0845homeinsurance.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
3789	www.0845householdinsurance.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A

ID	URL	Type	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α
3790	www.0845insurances.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
3791	www.0845lifecover.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
3893	www.360-credit-cards.co.uk	อินลิงก์	1	0	1	1	0	1	1	1	1	1	E
4127	www.all-car-insurances.co.uk	อินลิงก์	1	0	1	1	1	0	1	1	1	1	F
4128	www.all-carinsurance.co.uk	อินลิงก์	1	0	1	1	1	0	1	1	1	1	F
4129	www.all-carsinsurance.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
4130	www.all-carsinsurances.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
4138	www.allcar-insurance.co.uk	อินลิงก์	1	0	1	1	1	0	1	1	1	1	F
4139	www.allcar-insurances.co.uk	อินลิงก์	1	0	1	1	1	0	1	1	1	1	F
4141	www.allcars-insurances.co.uk	อินลิงก์	1	0	1	1	1	0	1	1	1	1	F
4142	www.allcarsinsurance.co.uk	อินลิงก์	1	0	1	1	1	0	1	1	1	1	F
4143	www.allcarsinsurances.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
4365	www.askliz.co.uk	อินลิงก์	0	0	0	0	0	0	1	0	0	0	G
5027	www.car-insurance-comparison.co.uk	อินลิงก์	1	0	1	1	1	1	1	1	1	1	A
5237	www.cheap-secured-loan.co.uk	อินลิงก์	1	0	0	1	1	1	1	1	1	1	H
5618	www.creditcardjunction.co.uk	อินลิงก์	1	0	0	1	0	1	1	1	1	1	C
5799	www.debt-consolidation-managment.co.uk	อินลิงก์	0	0	0	1	1	0	1	0	0	0	I
6821	www.get4me.co.uk	อินลิงก์	1	0	0	0	1	1	1	1	1	1	B
6882	www.go2net.co.uk	อินลิงก์	1	0	0	1	1	1	1	1	1	1	H
7449	www.insurance-guide.org.uk	อินลิงก์	1	0	0	1	1	1	1	1	1	1	H
8013	www.loanjunction.co.uk	อินลิงก์	0	0	0	1	0	1	1	1	1	1	J
8407	www.moneyjunction.co.uk	อินลิงก์	1	0	0	1	1	1	1	1	1	1	H
9709	www.search-today.co.uk	อินลิงก์	0	0	0	1	0	0	1	0	0	0	K
9711	www.search2find.co.uk	อินลิงก์	1	0	0	1	0	1	1	0	1	1	L
9712	www.search2go.co.uk	อินลิงก์	1	0	0	1	0	1	1	1	1	1	C
9716	www.search4more.co.uk	อินลิงก์	1	0	0	1	0	1	1	0	1	1	L
9735	www.searchtwice.co.uk	อินลิงก์	1	0	0	1	0	1	1	0	1	1	L
10441	www.themortgageauction.co.uk	อินลิงก์	1	0	1	1	0	1	1	1	1	1	M
10836	www.usearch.co.uk	อินลิงก์	0	0	1	1	0	0	1	0	0	0	N
11026	www.webbusinessopportunity.co.uk	อินลิงก์	1	0	0	1	0	1	1	1	1	1	C
11257	www.your-debt.co.uk	อินลิงก์	1	0	1	1	0	1	1	1	1	1	M
4140	www.allcarinsurance.co.uk	เอาท์ลิงก์	1	0	1	1	0	1	1	1	1	1	M
3792	www.0845motor.co.uk	ลิงก์สองทาง	1	0	1	1	0	1	1	1	1	1	M
3793	www.0845motorcycle.co.uk	ลิงก์สองทาง	1	0	1	1	0	1	1	1	1	1	M
3894	www.360-loan.co.uk	ลิงก์สองทาง	1	0	1	1	0	1	1	0	1	1	O

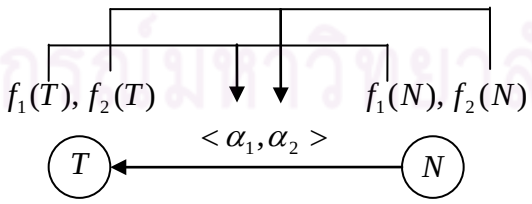
จากตารางที่ 3.1 สามารถหาลักษณะเฉพาะของเส้นเชื่อมได้ทั้งหมด 15 แบบ ซึ่งสามารถแสดงความถี่ของแต่ละลักษณะเฉพาะได้ตามตารางที่ 3.2

ตารางที่ 3.2 ตารางแสดงความถี่ของลักษณะเฉพาะของเส้นเชื่อมทั้ง 15 แบบ

ลักษณะเฉพาะของเส้นเชื่อม	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
ความถี่	17	2	5	1	1	6	1	4	1	1	1	3	5	1	1
เปอร์เซ็นต์	34	4	10	2	2	12	2	8	2	2	2	6	10	2	2

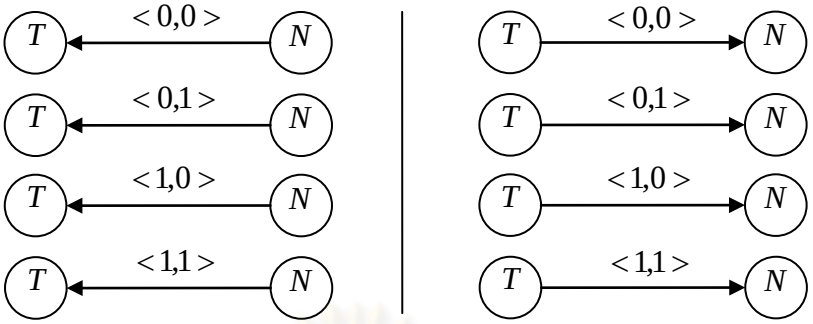
จากตารางที่ 3.2 จะเห็นได้ว่าลักษณะเฉพาะแบบ A มีจำนวนมากถึง 34% และยังมีลักษณะเฉพาะหลายชุดที่เกิดขึ้นเพียง 1 ครั้งเท่านั้น หากพิจารณาลักษณะเฉพาะที่มีความถี่โฮสต์ตั้งแต่สองโฮสต์ขึ้นไปจะประกอบไปด้วย 7 กลุ่ม คือ A, B, C, F, H, L และ M โดยแต่ละกลุ่มต่างก็ทำหน้าที่ชี้เข้าหาเว็บเป้าหมาย ซึ่งเปรียบเสมือนการเพิ่มคะแนนความสำคัญให้กับเว็บเป้าหมายโดยเฉพาะอย่างยิ่งมีโฮสต์ที่มีลักษณะเฉพาะของเส้นเชื่อมแบบ A □

ลักษณะเฉพาะดังกล่าว แสดงให้เห็นว่าลักษณะเฉพาะของเส้นเชื่อมระหว่างโฮสต์เป้าหมายกับโฮสต์เพื่อนบ้านโดยรอบ หากนำมาใช้เป็นป้ายกำกับเส้นเชื่อมในโฮสต์กราฟแล้ว จะเปรียบเสมือนการกำหนดป้ายกำกับให้กับเส้นเชื่อมโดยเส้นเชื่อมที่มีลักษณะเหมือนกัน ก็จะมีป้ายกำกับเดียวกัน สำหรับการทดลองนี้การสร้างไวยากรณ์กราฟโดยการอนุมาน เพื่อตรวจจับลิงก์ฟาร์ม จำเป็นต้องกำหนดลักษณะเฉพาะของลิงก์ฟาร์มที่แตกต่างออกจากเว็บเพจปกติทั่วไปก่อน ทั้งนี้เพื่อสามารถนำไวยากรณ์กราฟลิงก์ฟาร์มไปตรวจจับโครงสร้างลิงก์ฟาร์มอื่นได้อย่างมีประสิทธิภาพซึ่งการกำหนดลักษณะเฉพาะในงานวิจัยนี้จะใช้การสร้างป้ายกำกับเส้นเชื่อม เพื่อให้โครงสร้างของโฮสต์ปกติแยกออกจากโฮสต์สแปมได้อย่างมีประสิทธิภาพ ซึ่งในการทดลองนี้ได้ทำการวิเคราะห์โครงสร้างกราฟใน 2 ระดับ จากโหนดเป้าหมายและใช้ลักษณะเด่นในฟังก์ชันกำกับเส้นเชื่อม 2 ลักษณะเด่นคือ คะแนนเพจแรงค์ และคะแนนทรังเคทเพจแรงค์ ทั้งนี้เพราะคะแนนการจัดลำดับทั้งสองสามารถแสดงพฤติกรรมการถ่ายเทคะแนนการจัดอันดับได้ ดังนั้นฟังก์ชันการกำกับเส้นเชื่อมสามารถแสดงได้ดังรูปที่ 3.7



รูปที่ 3.7 ฟังก์ชันกำกับเส้นเชื่อมสำหรับ 2 ลักษณะเด่น

จากนั้นกำหนดป้ายกำกับเส้นเชื่อมด้วยฟังก์ชันกำกับเส้นเชื่อมดังกล่าวแล้วจะพบว่าโครงสร้างของโฮสต์กราฟจะมีรูปแบบของเส้นเชื่อมในระดับแรกเกิดขึ้น 8 รูปแบบเมื่อพิจารณาเส้นเชื่อมระหว่างโฮสต์เป้าหมายและโฮสต์เพื่อนบ้าน ดังรูปที่ 3.8



รูปที่ 3.8 รูปแบบของเส้นเชื่อมทั้ง 8 รูปแบบ

รูปแบบของเส้นเชื่อมทั้ง 8 แบบนี้ เป็นรูปแบบที่เป็นไปได้ทั้งหมดของเส้นเชื่อมระหว่างสองโหนดใดๆ ซึ่งในการอนุมานไวยากรณ์กราฟจะวิเคราะห์การเกิดรูปแบบของเส้นเชื่อมเหล่านี้โดยพิจารณาสัดส่วนการเกิดลักษณะเส้นเชื่อมในแต่ละแบบเพื่อใช้เป็นลักษณะเฉพาะในการแยกลิงก์ฟาร์มออกจากเว็บปกติต่อไป

3.3 การอนุมานไวยากรณ์กราฟและการตรวจจับลิงก์ฟาร์ม

ขั้นตอนวิธีการตรวจจับลิงก์ฟาร์มโดยวิธีการอนุมานไวยากรณ์กราฟ เริ่มต้นด้วยการสร้างป้ายกำกับให้กับเว็บกราฟด้วยการกำหนดป้ายกำกับเส้นเชื่อมจากนั้นแบ่งข้อมูลออกเป็นถังข้อมูล 10 ถังตามคะแนนเพจแรงค์ทั้งนี้เนื่องจากว่าลักษณะโครงสร้างของลิงก์ฟาร์มในช่วงคะแนนเพจแรงค์ที่ต่างกันจะมีโครงสร้างที่แตกต่างกันออกไปดังที่กล่าวมาแล้ว จากนั้นทำการแบ่งข้อมูลออกเป็นชุดข้อมูลสอน (training set) และชุดข้อมูลทดสอบ (test set) สำหรับชุดข้อมูลสอนนั้นจะนำมาผ่านกระบวนการอนุมานไวยากรณ์ เพื่อหาไวยากรณ์กราฟทั้งตัวอย่างบวก (โฮสต์สแปม) และตัวอย่างลบ (โฮสต์ปกติ) เพื่อใช้ในกระบวนการแจงส่วนสำหรับตรวจสอบกับชุดข้อมูลทดสอบต่อไป ทั้งนี้เมื่อผ่านกระบวนการอนุมานไวยากรณ์กราฟแล้วจะได้ไวยากรณ์กราฟแต่ละโปรดักชันพร้อมทั้งค่าเฉลี่ยในการใช้แต่ละโปรดักชัน และถูกกำหนดให้เป็นค่ากลาง (centroid) ซึ่งชุดข้อมูลสอนประกอบด้วยโฮสต์สแปมและโฮสต์ปกติ ดังนั้นค่ากลางจะประกอบไปด้วยค่ากลางของโฮสต์สแปมและโฮสต์ปกติดังนี้

ค่ากลางของโฮสต์สแปม (C_s) ประกอบไปด้วยค่าเฉลี่ยของเปอร์เซ็นต์การใช้โปรดักชันแต่ละโปรดักชันของชุดข้อมูลสอนที่เป็นโฮสต์สแปมสามารถเขียนเป็นเซตได้ดังนี้

$$C_s = \{C_{s,1}, C_{s,2}, C_{s,3}, ..., C_{s,m}\}$$

สำหรับค่ากลางของโฮสต์ปกติ (C_n) ประกอบด้วยค่าเฉลี่ยของเปอร์เซ็นต์การใช้โปรดักชันแต่ละโปรดักชันของชุดข้อมูลสอนที่เป็นโฮสต์ปกติสามารถเขียนเป็นเซตได้ดังนี้

$$C_n = \{C_{n,1}, C_{n,2}, C_{n,3}, ..., C_{n,m}\}$$

เซตของค่ากลางทั้งโฮสต์สแปมและโฮสต์ปกติสามารถคำนวณค่าเฉลี่ยของเปอร์เซ็นต์การใช้โปรดักชันได้ดังสมการ

$$C_{s,i} = \frac{\sum_{k=1}^{j} P_{i,k}}{j}$$

$$C_{n,i} = \frac{\sum_{k=1}^{j} P_{i,k}}{j}$$

เมื่อ $C_{s,i}$ คือค่าเฉลี่ยจำนวนการใช้โปรดักชันที่ i ของโฮสต์สแปม

$C_{n,i}$ คือค่าเฉลี่ยจำนวนการใช้โปรดักชันที่ i ของโฮสต์ปกติ

$P_{i,k}$ คือเปอร์เซ็นต์การใช้โปรดักชันที่ i ของโฮสต์ที่ k

i คือลำดับของโปรดักชันที่ i โดย i={1,2,3,…,m}

k คือลำดับของโฮสต์ที่ k โดย k={1,2,3,…,j}

ขั้นตอนต่อไปคือกระบวนการแจงส่วนเพื่อใช้ตรวจสอบโฮสต์สแปมจากชุดข้อมูลทดสอบ เมื่อโฮสต์แต่ละโฮสต์ในชุดข้อมูลทดสอบถูกแจงส่วน และให้ค่าเฉลี่ยในการใช้โปรดักชันแต่ละโปรดักชัน จากนั้นพิจารณาความห่าง (distant) ระหว่างเปอร์เซ็นต์ในการใช้โปรดักชันที่ได้จากโฮสต์ทดสอบกับค่ากลางของโฮสต์สแปมและโฮสต์ปกติ ซึ่งค่าความห่างสำหรับโฮสต์สแปมสามารถแสดงเป็นสมการได้ดังนี้

$$D_s = \sqrt{(p_1 - C_{s,1})^2 + (p_2 - C_{s,2})^2 + (p_3 - C_{s,3})^2 + ... + (p_n - C_{s,m})^2}$$

และค่าความห่างของโฮสต์ปกติสามารถแสดงได้สมการดังนี้

$$D_n = \sqrt{(p_1 - C_{n,1})^2 + (p_2 - C_{n,2})^2 + (p_3 - C_{n,3})^2 + ... + (p_n - C_{n,m})^2}$$

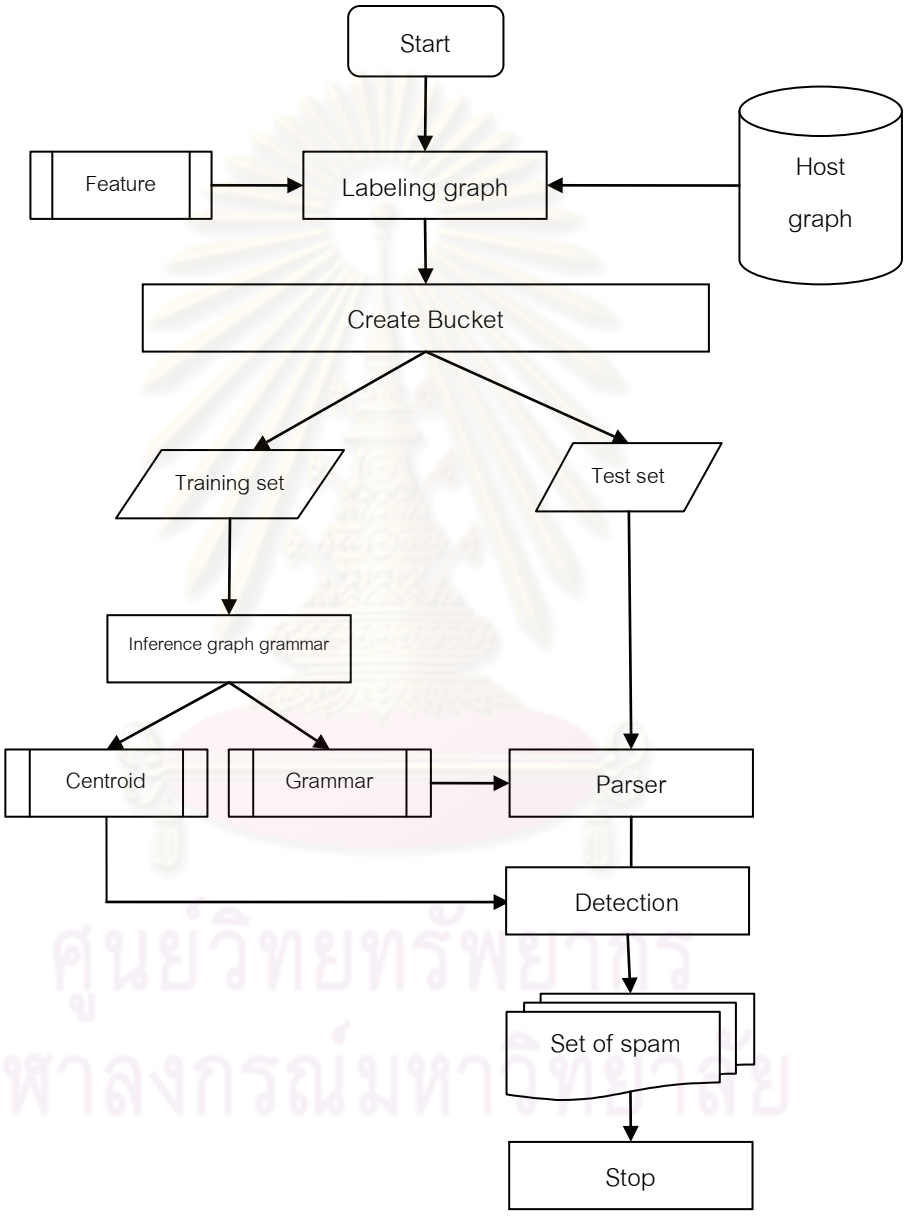
เมื่อ p_i คือเปอร์เซ็นต์การใช้โปรดักชันที่ i โดย i={1,2,3,…,n}

D_s คือความห่างจากค่ากลางชุดข้อมูลสอนโฮสต์สแปม

D_n คือความห่างจากค่ากลางชุดข้อมูลสอนโฮสต์ปกติ

ค่าความห่างนี้จะใช้ตัดสินว่าโฮสต์ทดสอบที่พิจารณาควรจะจัดอยู่ในกลุ่มของโฮสต์สแปมหรือไม่ โดยพิจารณาระยะทางความห่างของโฮสต์ที่ทดสอบกับค่ากลางทั้งค่ากลางของโฮสต์ปกติ

และค่ากลางของโฮสต์สแปม หากมีความความห่างของโฮสต์ปกติมากกว่าความห่างของโฮสต์สแปมจะถูกตัดสินว่าเป็นสแปม แต่หากมีความความห่างของโฮสต์สแปมมากกว่าความห่างของโฮสต์ปกติจะถูกตัดสินว่าเป็นปกติ ซึ่งกระบวนการทั้งหมดนี้สามารถแสดงเป็นผังงานได้ตามรูปที่ 3.9 ในส่วนต่อไปจะอธิบายถึงการทำงานของอัลกอริทึมการอนุมานไวยากรณ์กราฟและตรวจจับลิงก์ฟาร์มในแต่ละขั้นตอนอย่างละเอียด รวมทั้งอัลกอริทึมย่อยในการตรวจจับลิงก์ฟาร์มด้วย



รูปที่ 3.9 ผังงานแสดงอัลกอริทึมการอนุมานไวยากรณ์กราฟและการตรวจจับลิงก์ฟาร์ม

3.3.1 อัลกอริทึมการอนุมานไวยากรณ์กราฟและการตรวจจับลิงก์ฟาร์ม

อัลกอริทึมการอนุมานไวยากรณ์กราฟและการตรวจจับลิงก์ฟาร์ม ประกอบไปด้วยอัลกอริทึมย่อยคือ อัลกอริทึมการกำหนดป้ายกำกับเส้นเชื่อม อัลกอริทึมการแบ่งถังข้อมูล อัลกอริทึมการอนุมานไวยากรณ์กราฟ และอัลกอริทึมการแจงส่วนไวยากรณ์กราฟ ดังแสดงได้เป็นรหัสเทียมตามอัลกอริทึมที่ 3.1 ถึงอัลกอริทึมที่ 3.5

อัลกอริทึมที่ 3.1 อัลกอริทึมการอนุมานไวยากรณ์กราฟและการตรวจจับลิงก์ฟาร์ม

Algorithm 3.1: Link farm detection

LINK_FARM_DETECTION (*Host graph, Training set, Test set, Features, PageRank*)

1. **Begin**
2. *Features* = $\{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n\}$
3. EDGE_LABELING(*Host graph, Features*)
4. BUCKET(*Host graph, PageRank)*
5. INFERENCE(*Training set, Host graph*)
6. **return** *Grammars* and *Centroids*
7. PARSER(*Grammars, Centroids, Test set*)
8. **return** *set of spam*
9. **End**

จากอัลกอริทึมการอนุมานไวยากรณ์กราฟและการตรวจจับลิงก์ฟาร์ม ข้อมูลนำเข้าคือ โฮสต์กราฟ ชุดข้อมูลสอน ชุดข้อมูลทดสอบ ชุดข้อมูลลักษณะเด่น และคะแนนเพจแรงค์ โดยผ่านอัลกอริทึมย่อยแรกคืออัลกอริทึมกำหนดป้ายกำกับเส้นเชื่อมเพื่อกำหนดป้ายกำกับให้เส้นเชื่อมทุกเส้นบนโฮสต์กราฟ จากนั้นจะทำการแบ่งข้อมูลออกเป็น 10 ถังด้วยอัลกอริทึมแบ่งถังข้อมูล จึงนำเข้าสู่อัลกอริทึมการอนุมานไวยากรณ์กราฟเพื่อหาไวยากรณ์กราฟและค่าเฉลี่ยของเปอร์เซ็นต์การใช้โปรดักชันในแต่ละโปรดักชัน (ค่ากลาง) ในส่วนสุดท้ายคือการแจงส่วนกับชุดข้อมูลทดสอบด้วยอัลกอริทึมการแจงส่วน สามารถอธิบายขั้นตอนการทำงานอย่างละเอียดได้ดังนี้

บรรทัดที่ 2 : กำหนดข้อมูลลักษณะเด่น

บรรทัดที่ 3 : สร้างป้ายกำกับเส้นเชื่อมด้วยอัลกอริทึมกำหนดป้ายกำกับเส้นเชื่อม

บรรทัดที่ 4 : แบ่งถังข้อมูลออกเป็น 10 ถัง ด้วยอัลกอริทึมแบ่งถังข้อมูล

บรรทัดที่ 5 : อนุมานไวยากรณ์กราฟด้วยข้อมูลสอน

บรรทัดที่ 6 : ได้ไวยากรณ์กราฟจากการอนุมานพร้อมค่าเฉลี่ยกลางของแต่ละโปรดักชัน

บรรทัดที่ 7 : ตรวจสอบโฮสต์สแปมด้วยการแจงส่วนไวยากรณ์กราฟที่ได้จากการอนุมาน

บรรทัดที่ 8 : ให้ผลลัพธ์เป็นเซตของโฮสต์สแปม

3.3.2 อัลกอริทึมกำหนดป้ายกำกับเส้นเชื่อม

เนื่องจากเว็บกราฟ หรือโฮสต์กราฟเป็นกราฟที่ประกอบด้วยโหนดและเส้นเชื่อม ที่เชื่อมถึงกันอย่างหนาแน่น โดยแต่ละเส้นเชื่อมนั้นเป็นเส้นเชื่อมที่แสดงถึงการเชื่อมโยงกันเท่านั้น แต่สำหรับการอนุมานไวยากรณ์กราฟนั้นจะต้องใช้ป้ายกำกับเส้นเชื่อม ด้วยสมมติฐานที่ว่า เส้นเชื่อมที่มีหน้าที่ หรือมีลักษณะเหมือนกันจะมีป้ายกำกับเหมือนกัน ทั้งนี้เพื่อให้ได้ไวยากรณ์กราฟที่บ่งบอกถึงลักษณะโครงสร้างที่แท้จริงของโฮสต์สแปมและโฮสต์ปกติ ดังนั้นในงานวิจัยนี้จึงเสนออัลกอริทึมการกำหนดป้ายกำกับเส้นเชื่อม โดยใช้ลักษณะเด่นของโฮสต์ต้นทางและโฮสต์ปลายทางของเส้นเชื่อมใดๆ ดังแสดงตามรหัสเทียมอัลกอริทึมที่ 3.2

อัลกอริทึมที่ 3.2 อัลกอริทึมกำหนดป้ายกำกับเส้นเชื่อม

Algorithm 3.2: Edge labeling function

```
EDGE_LABELING (Host graph, Features)
1.  Begin
2.  repeat node in Host graph
3.      if found link around destination node
4.          repeat read Feature
5.              if Feature value of destination node ≤ Feature value of source node
6.                  label edge with “0”
7.              else
8.                  label edge with “1”
9.              end if
10.         until read all Features
11.     end if
12. until read all nodes in Host graph
13. End
```

อัลกอริทึมกำหนดป้ายกำกับเส้นเชื่อมสามารถอธิบายเป็นขั้นตอนโดยละเอียดได้ดังนี้

บรรทัดที่ 2 : อ่านข้อมูลโฮสต์กราฟแต่ละโฮสต์

บรรทัดที่ 3 : ตรวจสอบเส้นเชื่อมรอบๆ โฮสต์

บรรทัดที่ 4 : อ่านข้อมูลลักษณะเด่นแต่ละลักษณะ

บรรทัดที่ 5-6 : ถ้าค่าของลักษณะของโฮสต์ต้นทางน้อยกว่าหรือเท่ากับค่าของลักษณะเด่นของโฮสต์ปลายทางจะกำหนดป้ายกำกับเส้นเชื่อมด้วยค่า 0

บรรทัดที่ 7-8 : ถ้าค่าของลักษณะของโฮสต์ต้นทางมากกว่าค่าของลักษณะเด่นของโฮสต์ปลายทางจะกำหนดป้ายกำกับเส้นเชื่อมด้วยค่า 1

3.3.3 **อัลกอริทึมการแบ่งถังข้อมูล**

อัลกอริทึมสำหรับแบ่งถังข้อมูล ทำการแบ่งข้อมูลออกเป็น 10 ถังตามคะแนนเพจแรงค์ โดยถังที่ 1 จะเป็นถังที่หน้าเว็บเพจหลักของโฮสต์ดังกล่าวมีคะแนนเพจแรงค์สูงสุด และถังที่ 10 จะมีโฮสต์ที่มีคะแนนเพจแรงค์ต่ำสุดทั้งนี้เพราะโครงสร้างของลิงก์ฟาร์มในถังที่มีคะแนนเพจแรงค์สูงจะมีโครงสร้างแตกต่างออกไปกับโครงสร้างลิงก์ฟาร์มที่มีคะแนนเพจแรงค์ต่ำ เพื่อประสิทธิภาพในการตรวจจับที่ดียิ่งขึ้น สามารถแสดงเป็นรหัสเทียมได้ดังอัลกอริทึมที่ 3.3

อัลกอริทึมที่ 3.3 อัลกอริทึมการแบ่งถังข้อมูล

Algorithm 3.3: Divide bucket

BUCKET(*Host graph, PageRank)*

1. **Begin**
2. **repeat** host in *Host graph*
3. *Number of host* = Count all hosts in host graph
4. **until** read all of *Host graph*
5. $n = \lceil Number\ of\ host / 10 \rceil$
6. Sort host by *PageRank* score
7. **repeat** bucket start $i = 1$
8. $bucket[i] = \bigcup_{j=(i-1)n+1}^{i \times n} host[j]$
9. **until** $i = 10$
10. **End**

จากอัลกอริทึมแบ่งถังข้อมูลสามารถอธิบายเป็นขั้นตอนโดยละเอียดได้ดังนี้

บรรทัดที่ 2-4 : นับจำนวนโฮสต์ทั้งหมดบนเว็บกราฟ

บรรทัดที่ 5 : หาจำนวนโฮสต์ในแต่ละถัง ทั้ง 10 ถัง

บรรทัดที่ 6 : เรียงข้อมูลตามคะแนนเพจแรงค์

บรรทัดที่ 7-9 : แบ่งถังข้อมูลออกเป็น 10 ถัง ตามคะแนนเพจแรงค์

3.3.4 อัลกอริทึมการอนุมานไวยากรณ์กราฟลิงก์ฟาร์ม

การอนุมานไวยากรณ์กราฟลิงก์ฟาร์มด้วยเส้นเชื่อม จะประยุกต์การอนุมานจากอัลกอริทึมของจาเคค คุกลุก ที่ได้กล่าวไว้ในข้างต้น เพื่อหาไวยากรณ์กราฟจากโฮสต์สแปมบนโฮสต์กราฟ จากนั้นใช้การตรวจจับลิงก์ฟาร์มโดยการแจงส่วนไวยากรณ์กราฟเพื่อตรวจสอบความถูกต้องของไวยากรณ์กราฟที่ได้จากการอนุมาน ดังแสดงได้ตามอัลกอริทึมที่ 3.4

อัลกอริทึมที่ 3.4 อัลกอริทึมการอนุมานไวยากรณ์กราฟลิงก์ฟาร์ม

Algorithm 3.4: Inference graph grammar.

INFERENCE (*Training set*, *Host graph*)

1. **Begin**
2. *target* = {each host in *Training set*}
3. **repeat**
4. read *target*
5. initial *edge* = { }
6. *edge_level_1* = {all of edges around target node.}
7. **for each** edge $E_1 \in$ *edge_level_1*
8. *substructure* $\cup$ edge E_1
9. read neighbor node of each edge E_1.
10. *edge_level_2* = {all of edges around neighbor nodes level 1.}
11. **for each** edge $E_2 \in$ *edge_level_2*
12. substructure E = compose of edge E_1 and E_2
13. *substructure* $\cup E$
14. **end for**
15. initial *edge_level_2* = { }
16. **end for**
17. express substructure to graph grammar
18. **for each** production
19. *Centroid* = Average percent of number of each production
20. **end for**
21. **until** *target* = { }
22. **return** *grammar*
23. **End**

อัลกอริทึมการอนุมานไวยากรณ์กราฟลิงก์ฟาร์มด้วยเส้นเชื่อมสามารถอธิบายเป็นขั้นตอนโดยละเอียดได้ดังนี้

บรรทัดที่ 2 : นำโฮสต์ของชุดข้อมูลสอนแต่ละโฮสต์ที่เตรียมไว้เป็นโฮสต์เป้าหมายในการอนุมานไวยากรณ์กราฟ

บรรทัดที่ 4 : อ่านข้อมูลโฮสต์กราฟแต่ละโฮสต์จากชุดข้อมูลสอน

บรรทัดที่ 5 : กำหนดเซตของเส้นเชื่อมเริ่มต้นให้เป็นเซตว่าง

บรรทัดที่ 6 : กำหนดเซตของเส้นเชื่อมรอบโฮสต์เป้าหมายในระดับที่ 1

บรรทัดที่ 7 : พิจารณาเส้นเชื่อมแต่ละเส้นในเซตของเส้นเชื่อมระดับ 1

บรรทัดที่ 8 : เก็บรูปแบบของเส้นเชื่อมไว้ในเซตของโครงสร้างย่อย

บรรทัดที่ 9 : อ่านข้อมูลเส้นเชื่อมในระดับ 2 รอบๆ โฮสต์ของโหนดปลายทางจากเส้นเชื่อมระดับ 1

บรรทัดที่ 10 : กำหนดเซตของเส้นเชื่อมรอบโฮสต์เป้าหมายในระดับที่ 2

บรรทัดที่ 11 : พิจารณาเส้นเชื่อมแต่ละเส้นในเซตของเส้นเชื่อมระดับ 2

บรรทัดที่ 12 : เก็บรูปแบบของเส้นเชื่อมในระดับที่ 2 โดยนำมารวมกับระดับที่ 1

บรรทัดที่ 13 : เก็บรูปแบบของเส้นเชื่อมไว้ในเซตของโครงสร้างย่อย

บรรทัดที่ 15 : กำหนดค่าเริ่มต้นเซตของเส้นเชื่อมในระดับที่ 2 เพื่อทำงานในรอบต่อไป

บรรทัดที่ 17 : สร้างไวยากรณ์กราฟจากโครงสร้างย่อยที่พบในโครงสร้างโฮสต์ทดสอบ

บรรทัดที่ 18-20 : คำนวณค่ากลางโดยหาค่าเฉลี่ยของเปอร์เซ็นต์การใช้แต่ละโปรดักชัน

บรรทัดที่ 21 : ทำการสร้างไวยากรณ์กราฟจนกระทั่งครบทุกโฮสต์ของชุดข้อมูลสอน

บรรทัดที่ 22 : ได้ผลลัพธ์เป็นไวยากรณ์กราฟ

3.3.5 **อัลกอริทึมการแจงส่วนไวยากรณ์กราฟ**

การตรวจสอบประสิทธิภาพของไวยากรณ์กราฟที่ได้จากการอนุมานนั้นจะใช้อัลกอริทึมการแจงส่วนไวยากรณ์กราฟ โดยจะแจงส่วนของโครงสร้างโฮสต์ชุดข้อมูลทดสอบว่าตรงกับไวยากรณ์กราฟที่ได้จากอัลกอริทึมการอนุมานที่ได้กล่าวมาในข้างต้นหรือไม่ โดยจะพิจารณาจากโครงสร้างก่อนที่ตรงกันกับไวยากรณ์ จากนั้นจะคำนวณค่าความห่างระหว่างค่ากลางของโฮสต์สแปมกับค่ากลางของโฮสต์ปกติเพื่อตัดสินว่าควรจะจัดโฮสต์ทดสอบดังกล่าวว่าเป็นสแปมหรือไม่

อัลกอริทึมที่ 3.5 อัลกอริทึมการแจงส่วนไวยากรณ์กราฟ

Algorithm 3.5: Parser graph grammar.

PARSER (*Test set*, *Centroid*)

1. **Begin**
2. *target* = {each host in *Test set*}
3. **repeat**
4. read *target*
5. initial *edge* = { }
6. *edge_level_1* = {all of edges around target node.}
7. **for each** edge $E_1 \in$ *edge_level_1*
8. *substructure* $\cup$ edge E_1
9. read neighbor node of each edge E_1.
10. *edge_level_2* = {all of edges around neighbor node level 1.}
11. **for each** edge $E_2 \in$ *edge_level_2*
12. substructure E = compose of edge E_1 and E_2
13. *substructure* $\cup E$
14. **end for**
15. initial *edge_level_2* = { }
16. **end for**
17. **for each** production
18. D_n = distant from *Centroid* of normal
19. D_s = distant from *Centroid* of Spam
20. **if** (D_n - D_s > 0) Classify as "**Spam**"
21. **else** Classify as "**Normal**"
22. **end for**
23. **until** *target* = { }
24. **End**

อัลกอริทึมการแจงส่วนไวยากรณ์กราฟในช่วงแรกจะมีการทำงานคล้ายกับอัลกอริทึมการอนุมานไวยากรณ์กราฟ เนื่องจากต้องหาโครงสร้างย่อยที่ตรงกับไวยากรณ์กราฟ สามารถอธิบายเป็นขั้นตอนโดยละเอียดได้ดังนี้

บรรทัดที่ 2 : นำโฮสต์ของชุดข้อมูลสอนแต่ละโฮสต์ที่เตรียมไว้เป็นโฮสต์เป้าหมายในการอนุมานไวยากรณ์กราฟ

บรรทัดที่ 4 : อ่านข้อมูลโฮสต์กราฟแต่ละโฮสต์จากชุดข้อมูลสอน

บรรทัดที่ 5 : กำหนดเซตของเส้นเชื่อมเริ่มต้นให้เป็นเซตว่าง

บรรทัดที่ 6 : กำหนดเซตของเส้นเชื่อมรอบโฮสต์เป้าหมายในระดับที่ 1

บรรทัดที่ 7 : พิจารณาเส้นเชื่อมแต่ละเส้นในเซตของเส้นเชื่อมระดับ 1

บรรทัดที่ 8 : เก็บรูปแบบของเส้นเชื่อมไว้ในเซตของโครงสร้างย่อย

บรรทัดที่ 9 : อ่านข้อมูลเส้นเชื่อมในระดับ 2 รอบๆ โฮสต์ของโหนดปลายทางจากเส้นเชื่อมระดับ 1

บรรทัดที่ 10 : กำหนดเซตของเส้นเชื่อมรอบโฮสต์เป้าหมายในระดับที่ 2

บรรทัดที่ 11 : พิจารณาเส้นเชื่อมแต่ละเส้นในเซตของเส้นเชื่อมระดับ 2

บรรทัดที่ 12 : เก็บรูปแบบของเส้นเชื่อมในระดับที่ 2 โดยนำมารวมกับระดับที่ 1

บรรทัดที่ 13 : เก็บรูปแบบของเส้นเชื่อมไว้ในเซตของโครงสร้างย่อย

บรรทัดที่ 15 : กำหนดค่าเริ่มต้นเซตของเส้นเชื่อมในระดับที่ 2 เพื่อทำงานในรอบต่อไป

บรรทัดที่ 18 : คำนวนค่าความห่างจากค่ากลางของโฮสต์ปกติ

บรรทัดที่ 19 : คำนวนค่าความห่างจากค่ากลางของโฮสต์สแปม

บรรทัดที่ 20 : ถ้าผลต่างค่าความห่างของโฮสต์ปกติมีค่ามากกว่าค่าความห่างของโฮสต์สแปมจะตรวจจับว่าโฮสต์ดังกล่าวเป็นโฮสต์สแปม

บรรทัดที่ 21 : ถ้าผลต่างค่าความห่างของโฮสต์ปกติมีค่าน้อยกว่าหรือเท่ากับค่าความห่างของโฮสต์สแปมจะตรวจจับว่าโฮสต์ดังกล่าวเป็นโฮสต์ปกติ

บรรทัดที่ 23 : ทำการแจงส่วนไวยากรณ์กราฟจนกระทั่งครบทุกโฮสต์ของชุดข้อมูลทดสอบ

3.4 การวัดและทดสอบประสิทธิภาพการตรวจจับลิงก์ฟาร์ม

3.4.1 เครื่องมือสำหรับวัดประสิทธิภาพการทำงาน

การวัดผลความถูกต้องและประสิทธิภาพการทำงานของการตรวจจับลิงก์ฟาร์ม ในงานวิจัยทางด้านการตรวจจับเว็บสแปมหรือลิงก์ฟาร์ม จะใช้เครื่องมือในการวัดประสิทธิภาพในทางด้านการค้นคืนข้อมูล (information retrieval) ได้แก่ ค่าความแม่นยำ (precision) และค่าเรียกคืน (recall)

ค่าความแม่นยำ คือค่าที่ใช้วัดความถูกต้องของอัลกอริทึมในการตรวจจับโฮสต์สแปมที่เป็นสแปมจริง ว่ามีความถูกต้องเท่าไรเมื่อเทียบกับจำนวนโฮสต์ที่ถูกตรวจพบว่าเป็นสแปมโดยอัลกอริทึม

$$\text{ค่าความแม่นยำ} = \frac{\text{จำนวนโฮสต์สแปมที่ตรวจว่าเป็นสแปมจริง}}{\text{จำนวนโฮสต์ทั้งหมดที่ตรวจว่าเป็นสแปมโดยอัลกอริทึม}}$$

ค่าเรียกคืน คือค่าที่ใช้วัดประสิทธิการค้นคืนของอัลกอริทึมในการตรวจจับว่าสามารถตรวจจับโฮสต์สแปมได้มากน้อยเพียงใดเมื่อเทียบกับจำนวนโฮสต์สแปมทั้งหมดจากชุดข้อมูลทดสอบ

$$\text{ค่าเรียกคืน} = \frac{\text{จำนวนโฮสต์สแปมที่ตรวจว่าเป็นสแปมจริง}}{\text{จำนวนโฮสต์ทั้งหมดที่เป็นสแปมจากชุดข้อมูลทดสอบ}}$$

3.4.2 วิธีการวัดประสิทธิภาพการทำงานของอัลกอริทึม

การวัดประสิทธิภาพการทำงานของอัลกอริทึมตรวจจับลิงก์ฟาร์มโดยใช้การอนุมานไวยากรณ์กราฟนั้น ในการทดลองนี้สนใจการตรวจจับลิงก์ฟาร์มในกลุ่มที่มีคะแนนเพจแรงค์สูง 5 ถังแรก เนื่องจากโดยทั่วไปแล้ว เว็บเพจที่มีคะแนนเพจแรงค์สูง มักจะปรากฏอยู่ในหน้าแรกๆ หรืออันดับต้นของการค้นคืน ซึ่งผู้ใช้เองก็มักจะสนใจเว็บเพจที่ปรากฏอยู่ในลำดับต้นๆ เช่นกัน ดังนั้นการกำจัดหน้าเว็บที่เป็นสแปมออกจากผลการค้นคืนในลำดับต้นๆ ได้ จะทำให้ประสิทธิภาพการค้นคืนของโปรแกรมค้นหามีประสิทธิภาพและมีคุณภาพยิ่งขึ้น

ในการทดลองนี้จะการตรวจจับสแปมออกเป็น 10 ถัง ตามคะแนนเพจแรงค์ และใช้การตรวจสอบแบบไขว้ทั้งหมด 10 รอบการทำงาน (10-Fold cross validation) ด้วยสัดส่วนชุดข้อมูลสอนที่แตกต่างกัน 5 ระดับ คือ 10% 20% 30% 40% และ 50% ตามลำดับ จากชุดข้อมูลสุ่มที่ติดฉลากไว้แล้ว ในแต่ละระดับสัดส่วนข้อมูลสอน สามารถแสดงชุดข้อมูลสอนและชุดข้อมูลทดสอบได้ดังตารางที่ 3.3 ถึง ตารางที่ 3.7

ตารางที่ 3.3 ชุดข้อมูลการตรวจสอบแบบไขว้ 10 รอบ

รอบ	ชุดข้อมูลทดสอบ				
	สัดส่วน ชุดข้อมูลสอน 10%	สัดส่วน ชุดข้อมูลสอน 20%	สัดส่วน ชุดข้อมูลสอน 30%	สัดส่วน ชุดข้อมูลสอน 40%	สัดส่วน ชุดข้อมูลสอน 50%
1	2,3,4,5,6,7,8,9,10	3,4,5,6,7,8,9,10	4,5,6,7,8,9,10	5,6,7,8,9,10	6,7,8,9,10
2	3,4,5,6,7,8,9,10,1	4,5,6,7,8,9,10,1	5,6,7,8,9,10,1	6,7,8,9,10,1	7,8,9,10,1
3	4,5,6,7,8,9,10,1,2	5,6,7,8,9,10,1,2	6,7,8,9,10,1,2	7,8,9,10,1,2	8,9,10,1,2
4	5,6,7,8,9,10,1,2,3	6,7,8,9,10,1,2,3	7,8,9,10,1,2,3	8,9,10,1,2,3	9,10,1,2,3
5	6,7,8,9,10,1,2,3,4	7,8,9,10,1,2,3,4	8,9,10,1,2,3,4	9,10,1,2,3,4	10,1,2,3,4
6	7,8,9,10,1,2,3,4,5	8,9,10,1,2,3,4,5	9,10,1,2,3,4,5	10,1,2,3,4,5	1,2,3,4,5
7	8,9,10,1,2,3,4,5,6	9,10,1,2,3,4,5,6	10,1,2,3,4,5,6	1,2,3,4,5,6	2,3,4,5,6
8	9,10,1,2,3,4,5,6,7	10,1,2,3,4,5,6,7	1,2,3,4,5,6,7	2,3,4,5,6,7	3,4,5,6,7
9	10,1,2,3,4,5,6,7,8	1,2,3,4,5,6,7,8	2,3,4,5,6,7,8	3,4,5,6,7,8	4,5,6,7,8
10	1,2,3,4,5,6,7,8,9	2,3,4,5,6,7,8,9	3,4,5,6,7,8,9	4,5,6,7,8,9	5,6,7,8,9

3.4.3 การวิเคราะห์ประสิทธิภาพการทำงานของอัลกอริทึม

การวัดประสิทธิภาพการทำงานในการทดลองนี้จะทำการวัดประสิทธิภาพการทำงานทั้งในกลุ่มของโฮสต์ที่มีคะแนนเพจแรงค์สูง 5 ถังแรก และทั้ง 10 ถังข้อมูล เพื่อทำการเปรียบเทียบประสิทธิภาพการทำงานของการตรวจจับด้วยไวยากรณ์กราฟลิงก์ฟาร์มจากการอนุมานกับงานวิจัยที่ใช้ฐานข้อมูลทดสอบเดียวกัน ได้แก่ งานวิจัยการตรวจจับลิงก์ฟาร์มโดยไวยากรณ์กราฟที่ได้จากแบบจำลองลิงก์ฟาร์มที่เหมาะสม [16] งานวิจัยการตรวจจับเว็บสแปมโดยใช้วิธีการแอนตี้ทรัสต์ [17] และงานวิจัยการตรวจจับลิงก์สแปมแบบทรานดักทิฟ [18]

บทที่ 4

ผลการทดลอง

ในผลการทดลองนี้จะอธิบายถึงชุดข้อมูล รูปแบบเส้นเชื่อมลักษณะต่างๆ สำหรับสร้างไวยากรณ์กราฟ จากนั้นจะนำเสนอไวยากรณ์กราฟสำหรับตรวจจับลิงก์ฟาร์ม และนำเสนอประสิทธิภาพการตรวจจับลิงก์ฟาร์มในถังที่มีคะแนนเพจแรงค์สูง 5 ถังแรก และทำการเปรียบเทียบผลการทดลองกับงานวิจัยที่เกี่ยวข้อง [16-18] ซึ่งใช้ชุดข้อมูลทดสอบบนเว็บกราฟเดียวกัน ทั้งนี้จะวัดประสิทธิภาพของสัดส่วนข้อมูลสอน ค่าความแม่นยำ และค่าเรียกคืน

4.1 ชุดข้อมูล

ชุดข้อมูลในงานวิจัยนี้จะใช้ชุดข้อมูลที่มีนักวิจัยในด้านการตรวจจับเว็บสแปมนิยมใช้อย่างแพร่หลาย รวบรวมโดยศูนย์วิจัยยาฮู (Yahoo Research) ในปี ค.ศ.2006 [11] ประกอบไปด้วยชุดข้อมูลโดเมนประเทศอังกฤษ (.uk) รวมทั้งสิ้น 11,402 โฮสต์ หรือประมาณ 77 ล้านเว็บเพจ โดยโฮสต์ดังกล่าวได้ถูกจำแนกประเภทไว้โดยทีมนักวิจัยอาสาสมัคร สามารถแบ่งข้อมูลออกเป็น 3 กลุ่ม ได้แก่ โฮสต์ปกติ (normal host) จำนวน 4,948 โฮสต์ โฮสต์สแปม (spam host) จำนวน 674 โฮสต์ และโฮสต์ที่ไม่ได้ตัดสิน (undecided host) จำนวน 5,780 โฮสต์ ดังตารางที่ 4.1

ตารางที่ 4.1 จำนวนโฮสต์ในแต่ละประเภท

ประเภทโฮสต์	จำนวน (โฮสต์)	เปอร์เซ็นต์ (%)
โฮสต์ปกติ	4,948	43.40
โฮสต์สแปม	674	5.91
โฮสต์ที่ยังไม่ได้ตัดสิน	5,780	50.69
	11,402	

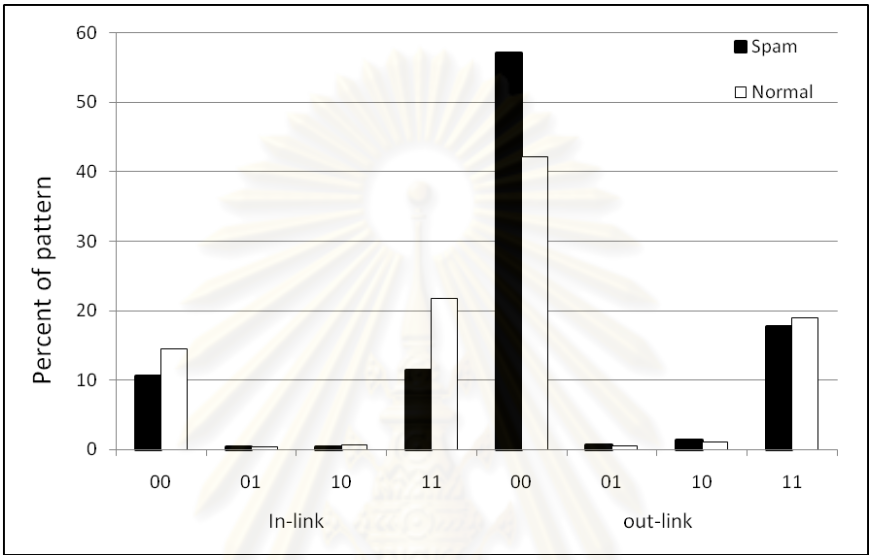
โดยในงานวิจัยนี้จะใช้โฮสต์กราฟในกระบวนการอนุมานไวยากรณ์กราฟ และจะไม่นำชุดโฮสต์ที่ยังไม่ได้ตัดสินมาพิจารณา

4.2 การอนุมานไวยากรณ์กราฟ

ในส่วนการทดลองการอนุมานไวยากรณ์กราฟจะใช้ชุดข้อมูลทดสอบสำหรับสร้างไวยากรณ์กราฟทั้งโฮสต์สแปมที่เป็นตัวอย่างบวก และโฮสต์ปกติที่เป็นตัวอย่างลบ เมื่อนำชุดข้อมูลทดสอบผ่านอัลกอริทึมการอนุมานไวยากรณ์กราฟแล้วจะได้โปรดักชันของไวยากรณ์กราฟรูปแบบต่างๆ จำนวนมาก แต่การทดลองนี้จะวิเคราะห์สัดส่วนการใช้โปรดักชันแต่ละโปรดักชันรอบโฮสต์เป้าหมายที่พิจารณา และเลือกโปรดักชันของไวยากรณ์กราฟที่มีความแตกต่างของสัดส่วนระหว่าง

โฮสต์สแปมและโฮสต์ปกติ ซึ่งสามารถวิเคราะห์การเลือกโปรดักชันได้จากสัดส่วนการเกิดรูปแบบดังนี้

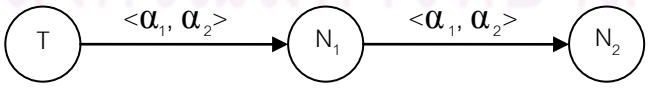
การอนุมานไวยากรณ์กราฟระดับ 1 จากโฮสต์เป้าหมาย สามารถพิจารณารูปแบบของโปรดักชันที่เกิดขึ้นในโครงสร้างระดับแรก ซึ่งจะเกิดรูปแบบทั้งหมด 8 รูปแบบโดยประกอบไปด้วยอินลิงก์ 4 แบบ และเอาท์ลิงก์ 4 แบบ เรียงตามลำดับดังรูปที่ 4.1



รูปที่ 4.1 เปอร์เซ็นต์จำนวนรูปแบบของเส้นเชื่อมในระดับ 1

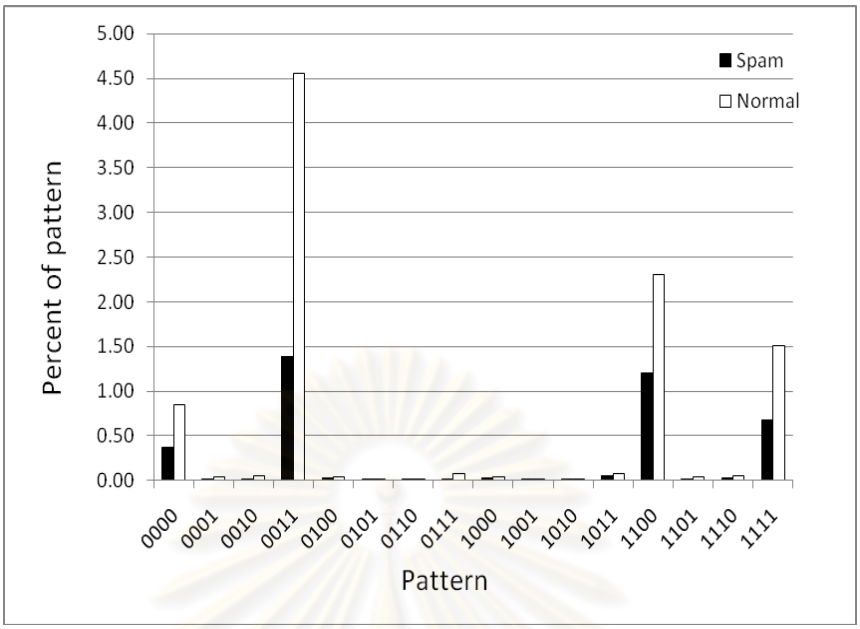
สำหรับการวิเคราะห์จำนวนการเกิดรูปแบบโครงสร้างในระดับที่ 2 นั้น จะเกิดรูปแบบของโครงสร้างทั้งหมด 64 รูปแบบ แบ่งออกเป็น 4 ชุด ดังนี้

ชุดที่ 1 ประกอบด้วย 16 รูปแบบ คือลิงก์แบบเอาท์ลิงก์ในระดับ 1 และลิงก์แบบเอาท์ลิงก์ในระดับ 2 ซึ่งป้ายกำกับบนเส้นเชื่อมสามารถแทนค่าด้วย 0 หรือ 1 ในแต่ละลักษณะเด่นดังแสดงในรูปที่ 4.2



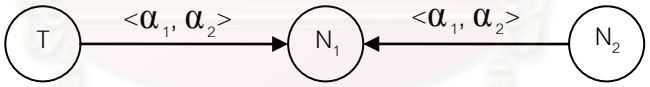
รูปที่ 4.2 รูปแบบโครงสร้างชุดที่ 1

จากนั้นพิจารณาลักษณะโครงสร้างลิงก์ โดยวิเคราะห์ค่าเฉลี่ยสัดส่วนของปริมาณรูปแบบลิงก์รอบๆ โฮสต์เป้าหมายทั้ง 16 รูปแบบสามารถเขียนเป็นกราฟได้ดังรูปที่ 4.3



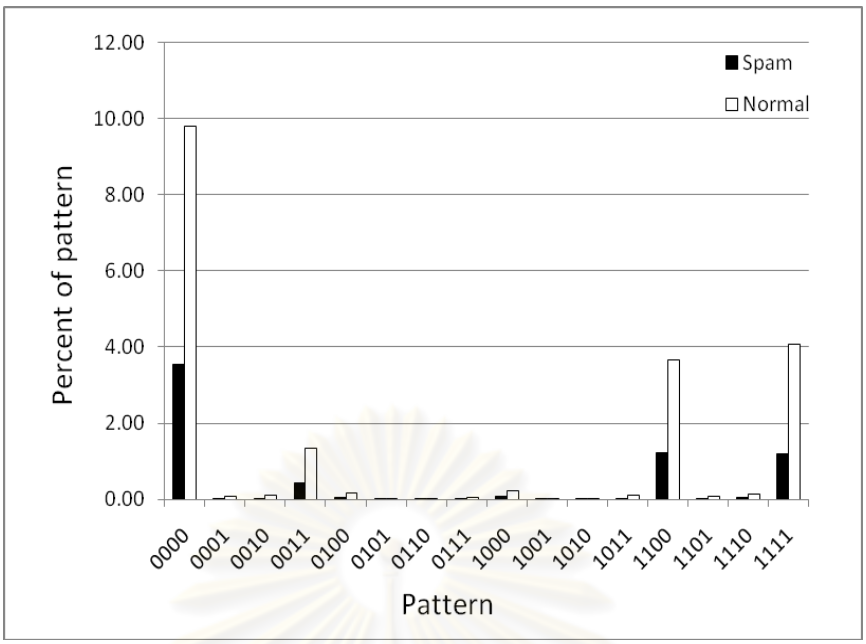
รูปที่ 4.3 เปอร์เซ็นต์จำนวนรูปแบบของเส้นเชื่อมใน 2 ระดับ ชุดที่ 1

ชุดที่ 2 ประกอบด้วย 16 รูปแบบ คือลิงก์แบบเอาท์ลิงก์ในระดับ 1 และลิงก์แบบอินลิงก์ในระดับ 2 ซึ่งป้ายกำกับบนเส้นเชื่อมสามารถแทนค่าด้วย 0 หรือ 1 ในแต่ละลักษณะเด่นดังแสดงในรูปที่ 4.4



รูปที่ 4.4 รูปแบบโครงสร้างชุดที่ 2

จากนั้นพิจารณาลักษณะโครงสร้างลิงก์ โดยวิเคราะห์ค่าเฉลี่ยสัดส่วนของปริมาณรูปแบบลิงก์รอบๆ โฮสต์เป้าหมายทั้ง 16 รูปแบบสามารถเขียนเป็นกราฟได้ดังรูปที่ 4.5

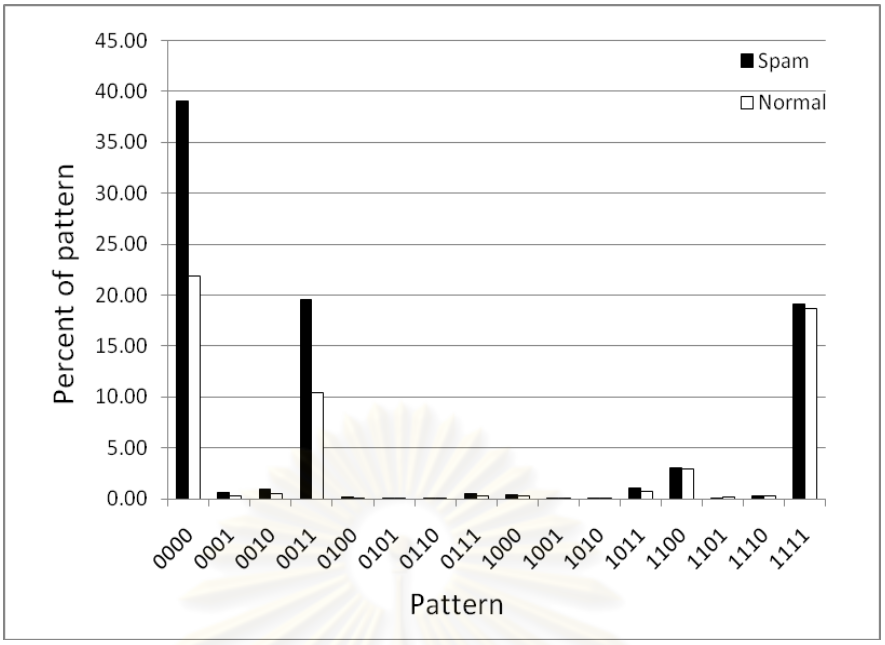


รูปที่ 4.5 เปอร์เซ็นต์จำนวนรูปแบบของเส้นเชื่อมใน 2 ระดับ ชุดที่ 2

ชุดที่ 3 ประกอบด้วย 16 รูปแบบ คือลิงก์แบบอินลิงก์ในระดับ 1 และลิงก์แบบเอาท์ลิงก์ในระดับ 2 ซึ่งป้ายกำกับบนเส้นเชื่อมสามารถแทนค่าด้วย 0 หรือ 1 ในแต่ละลักษณะเด่นดังแสดงในรูปที่ 4.6

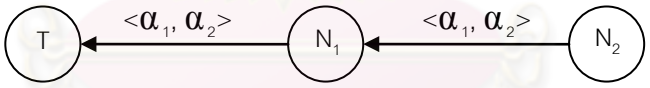
รูปที่ 4.6 รูปแบบโครงสร้างชุดที่ 3

จากนั้นพิจารณาลักษณะโครงสร้างลิงก์ โดยวิเคราะห์ค่าเฉลี่ยสัดส่วนของปริมาณรูปแบบลิงก์รอบๆ โฮสต์เป้าหมายทั้ง 16 รูปแบบสามารถเขียนเป็นกราฟได้ดังรูปที่ 4.7



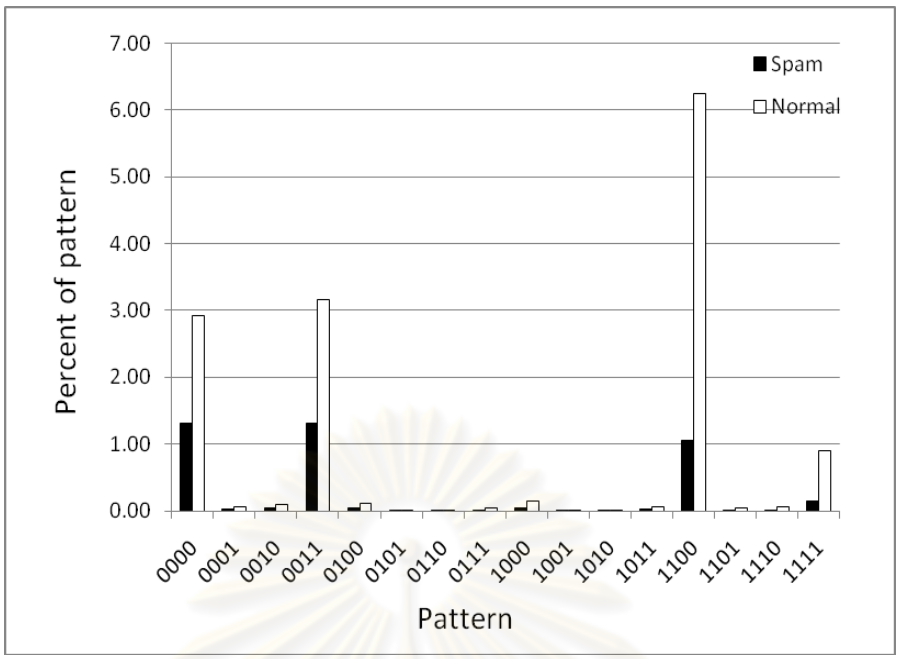
รูปที่ 4.7 เปอร์เซ็นต์จำนวนรูปแบบของเส้นเชื่อมใน 2 ระดับ ชุดที่ 3

ชุดที่ 4 ประกอบด้วย 16 รูปแบบ คือลิงก์แบบอินลิงก์ในระดับ 1 และลิงก์แบบอินลิงก์ในระดับ 2 ซึ่งป้ายกำกับบนเส้นเชื่อมสามารถแทนค่าด้วย 0 หรือ 1 ในแต่ละลักษณะเด่นดังแสดงในรูปที่ 4.8



รูปที่ 4.8 รูปแบบโครงสร้างชุดที่ 4

จากนั้นพิจารณาลักษณะโครงสร้างลิงก์ โดยวิเคราะห์ค่าเฉลี่ยสัดส่วนของปริมาณรูปแบบลิงก์รอบๆ โฮสต์เป้าหมายทั้ง 16 รูปแบบสามารถเขียนเป็นกราฟได้ดังรูปที่ 4.9

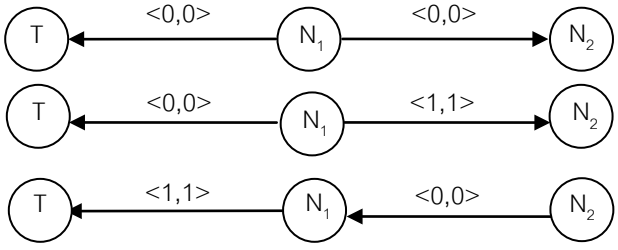


รูปที่ 4.9 เปอร์เซ็นต์จำนวนรูปแบบของเส้นเชื่อมใน 2 ระดับ ชุดที่ 4

จากรูปแบบของเส้นเชื่อมทั้ง 64 รูปแบบพบว่า ค่าเฉลี่ยของรูปแบบที่แสดงถึงความแตกต่างของลักษณะเฉพาะระหว่างลิงก์ฟาร์มมีอยู่หลายรูปแบบ ดังนั้นในการทดลองจะเลือกรูปแบบที่มีความแตกต่างอย่างเห็นได้ชัด รวมทั้งนำค่าส่วนเบี่ยงเบนมาตรฐานในแต่ละรูปแบบมาพิจารณาด้วย และได้ทำการคัดเลือกรูปแบบที่เหมาะสมสำหรับการใช้เป็นโปรดักชันของไวยากรณ์กราฟดังนี้

1. รูปแบบที่ 1: รูปแบบ 0000 ของชุดที่ 3 คือลิงก์แบบอินลิงก์ในระดับที่ 1 และลิงก์แบบเอาท์ลิงก์ในระดับที่ 2
2. รูปแบบที่ 2: รูปแบบ 0011 ของชุดที่ 3 คือลิงก์แบบอินลิงก์ในระดับที่ 1 และลิงก์แบบเอาท์ลิงก์ในระดับที่ 2
3. รูปแบบที่ 3: รูปแบบ 1100 ของชุดที่ 4 คือลิงก์แบบอินลิงก์ในระดับที่ 1 และลิงก์แบบอินลิงก์ในระดับที่ 2

รูปแบบเส้นเชื่อมทั้งสามรูปแบบสามารถแสดงโครงสร้างได้ดังรูปที่ 4.10 ตามลำดับ

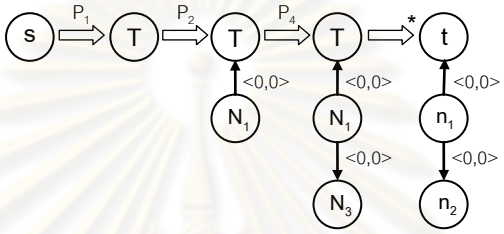


รูปที่ 4.10 รูปแบบโครงสร้างสำหรับสร้างไวยากรณ์กราฟทั้ง 3 รูปแบบ

4.3 ไวยากรณ์กราฟสำหรับตรวจจับลิงก์ฟาร์ม

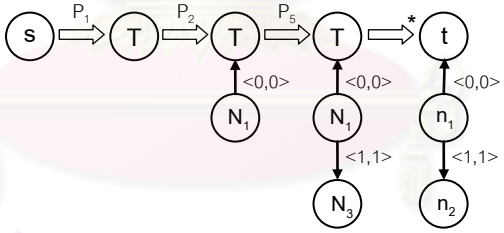
วิธีการสร้างไวยากรณ์กราฟสำหรับตรวจจับลิงก์ฟาร์มในการทดลองนี้ มีกระบวนการพิจารณาจากโครงสร้างย่อย (sub structure) ที่ซ้ำกันของโครงสร้างลิงก์ฟาร์ม สามารถแสดงได้ดังรูปที่ 4.14 ซึ่งไวยากรณ์กราฟที่ได้จากการอนุมานนี้สามารถใช้กฎการแปลงให้อยู่ในรูปแบบโครงสร้างทั้ง 3 รูปแบบ คือ

รูปแบบที่ 1: โครงสร้างรูปแบบที่ 1 สามารถสร้างด้วยโปรดักชันคือ $P_1 \rightarrow P_2 \rightarrow P_4 \rightarrow P_{10} \rightarrow P_8 \rightarrow P_7$ สามารถเขียนเป็นลำดับการแปลงได้ดังรูปที่ 4.11



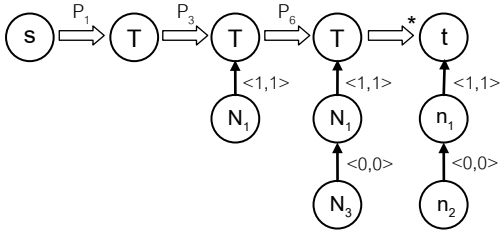
รูปที่ 4.11 ลำดับการแปลงของไวยากรณ์กราฟสำหรับโครงสร้างรูปแบบที่ 1

รูปแบบที่ 2: โครงสร้างรูปแบบที่ 2 สามารถสร้างด้วยโปรดักชันคือ $P_1 \rightarrow P_2 \rightarrow P_5 \rightarrow P_{11} \rightarrow P_8 \rightarrow P_7$ สามารถเขียนเป็นลำดับการแปลงได้ดังรูปที่ 4.12

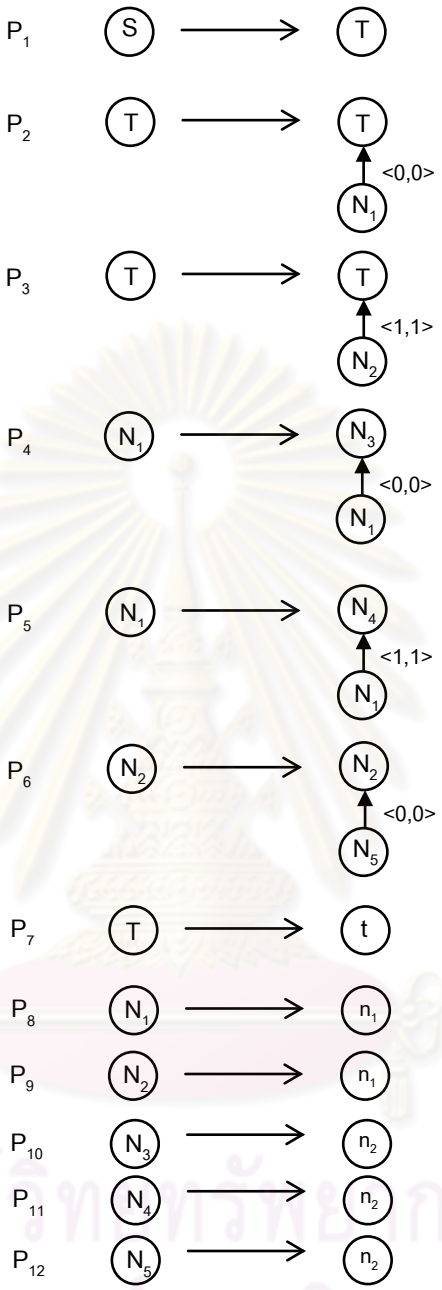


รูปที่ 4.12 ลำดับการแปลงของไวยากรณ์กราฟสำหรับโครงสร้างรูปแบบที่ 2

รูปแบบที่ 3: โครงสร้างรูปแบบที่ 3 สามารถสร้างด้วยโปรดักชันคือ $P_1 \rightarrow P_3 \rightarrow P_6 \rightarrow P_{12} \rightarrow P_9 \rightarrow P_7$ สามารถเขียนเป็นลำดับการแปลงได้ดังรูปที่ 4.13



รูปที่ 4.13 ลำดับการแปลงของไวยากรณ์กราฟสำหรับโครงสร้างรูปแบบที่ 3



โดย S คือ โหนดเริ่มต้น (start node)

T คือ โหนดเป้าหมายที่พิจารณา

N_1 และ N_2 คือ โหนดเพื่อนบ้าน ระยะทาง 1 จากโหนดเป้าหมาย

N_3 N_4และ N_5 คือ โหนดเพื่อนบ้าน ระยะทาง 2 จากโหนดเป้าหมาย

t, n_1,n_2 คือ โหนดปลายทาง (terminal node)

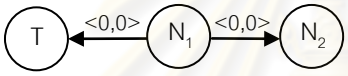
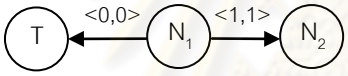
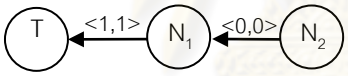
รูปที่ 4.14 ไวยากรณ์กราฟลิงก์ฟาร์มจากการอนุมานไวยากรณ์กราฟ

ในส่วนต่อไปแสดงตัวอย่างโครงสร้างโฮสต์สแปมและโฮสต์ปกติจะสร้างโดยใช้ไวยากรณ์กราฟดังที่กล่าวมาในข้างต้น ซึ่งในตัวอย่างนี้จะใช้ชุดข้อมูลสอนเพื่อหารูปแบบของโฮสต์สแปมและโฮสต์ปกติทั้งสองรูปแบบและจะสนใจรูปแบบโครงสร้างที่มีความแตกต่างกันของจำนวนการใช้โปรดักชันระหว่างโฮสต์สแปมและโฮสต์ปกติทั้ง 3 รูปแบบ

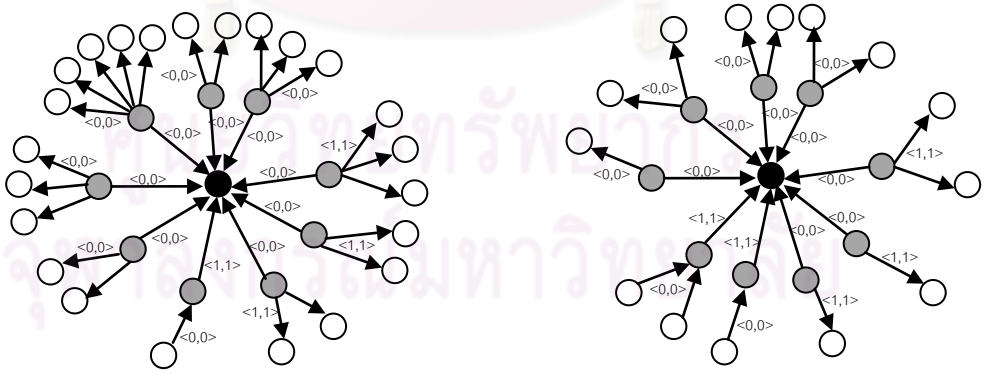
ตัวอย่างที่ 4.1 โครงสร้างโฮสต์สแปมและโฮสต์ปกติจากเปอร์เซ็นต์การใช้รูปแบบทั้ง 3 รูปแบบ

รูปแบบที่แสดงถึงลักษณะเฉพาะที่แตกต่างกันระหว่างโฮสต์สแปมและโฮสต์ปกติประกอบไปด้วย 3 รูปแบบ สามารถแสดงเปอร์เซ็นต์การใช้ในแต่ละรูปแบบได้ดังตารางที่ 4.2

ตารางที่ 4.2 เปอร์เซ็นต์การใช้รูปแบบในแต่ละโครงสร้าง

รูปแบบ	โครงสร้าง	เปอร์เซ็นต์การใช้รูปแบบ(%)	เปอร์เซ็นต์การใช้รูปแบบ(%)
1		24	39
2		11	19
3		6	1

จากตารางที่ 4.2 สามารถนำเปอร์เซ็นต์การใช้รูปแบบแต่ละรูปแบบมาสร้างเป็นโครงสร้างสำหรับโฮสต์สแปมและโฮสต์ปกติได้ดังรูปที่ 4.15



(a) โครงสร้างโฮสต์สแปม (b) โครงสร้างโฮสต์ปกติ

รูปที่ 4.15 โครงสร้างโฮสต์สแปมและโฮสต์ปกติ

จากรูปที่ 4.15 พบว่าโครงสร้างของโฮสต์สแปมมีความหนาแน่นของลิงก์โดยรอบโฮสต์เป้าหมายมากกว่าโฮสต์ปกติซึ่งสอดคล้องกับแบบจำลองลิงก์ฟาร์ม □

4.4 ผลการตรวจจับลิงก์ฟาร์มในระดับโฮสต์

การตรวจจับลิงก์ฟาร์มในระดับโฮสต์เมื่อนำชุดข้อมูลสอนผ่านกระบวนการอนุมานไวยากรณ์กราฟแล้วจะได้ไวยากรณ์กราฟพร้อมด้วยโปรดักชันต่างๆ สำหรับใช้ในการตรวจจับลิงก์ฟาร์ม การทดลองจะทำการตรวจสอบการตรวจจับโดยแบ่งถังข้อมูลออกเป็น 10 ถังข้อมูลตามคะแนนเพจแรงค์ แต่ในการทดลองนี้จะมุ่งตรวจจับลิงก์ฟาร์มกลุ่มที่มีคะแนนเพจแรงค์สูง 5 ถังแรก โดยใช้วิธีตรวจสอบแบบไขว้ 10 รอบในแต่ละถังข้อมูล และแบ่งชุดข้อมูลสอนในแต่ละรอบการทำงานเป็น 10% 20% 30% 40% และ 50% ตามลำดับ ผลการทดลองการตรวจวัดด้วยค่าความแม่นยำแสดงดังตารางที่ 4.3

ตารางที่ 4.3 ค่าความแม่นยำแต่ละช่วงสัดส่วนข้อมูลสอนในแต่ละถังข้อมูล

ถังที่	สัดส่วนข้อมูลสอน				
	10	20	30	40	50
1	74.8449	74.6897	75.1020	75.0349	75.6027
2	79.6364	79.0288	79.5173	79.6443	80.0449
3	84.3535	84.8340	85.1376	85.5800	85.6913
4	79.3566	84.5048	85.0169	85.6232	85.5581
5	64.1389	72.0002	78.0025	77.9610	78.7572
ค่าเฉลี่ย	76.4661	79.0115	80.5553	80.7687	81.1308

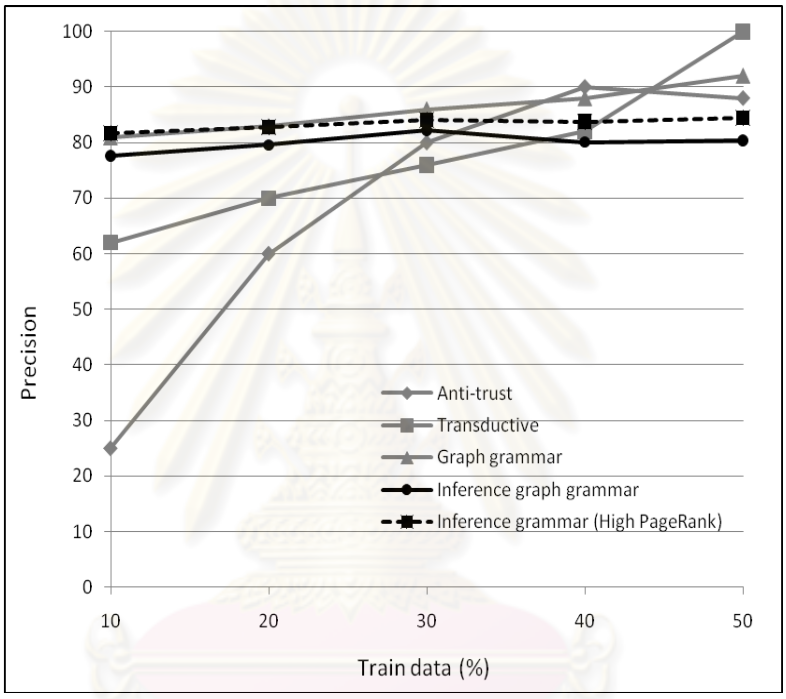
จากนั้นคำนวนค่าเรียกคืนของแต่ละช่วงสัดส่วนข้อมูลสอนในแต่ละถังข้อมูลที่ 1 ถึงถังข้อมูลที่ 5 สามารถแสดงค่าเรียกคืนได้ดังตารางที่ 4.4

ตารางที่ 4.4 ค่าเรียกคืนแต่ละช่วงสัดส่วนข้อมูลสอนในแต่ละถังข้อมูล

ถังที่	สัดส่วนข้อมูลสอน				
	10	20	30	40	50
1	81.8889	83.0000	82.8571	83.0000	83.0000
2	76.5079	75.8929	76.3265	75.7143	75.7143
3	80.9524	81.4286	81.8367	82.8571	82.8571
4	82.0370	82.2917	85.0169	82.2222	82.3333
5	56.1111	64.6875	78.0025	59.6429	72.0000
ค่าเฉลี่ย	75.49946	77.46014	80.80794	76.6873	79.18094

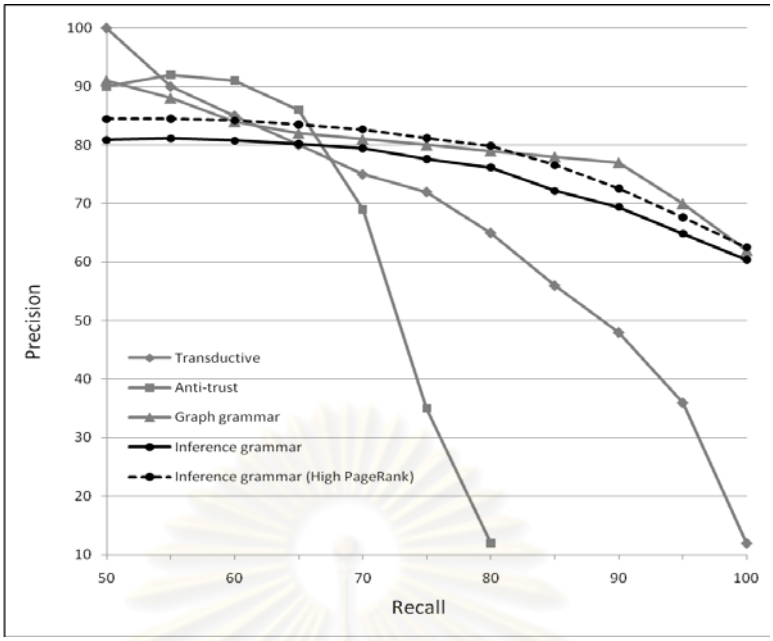
4.5 ผลการเปรียบเทียบประสิทธิภาพการตรวจจับลิงก์ฟาร์ม

การเปรียบเทียบประสิทธิภาพอัลกอริทึมการตรวจจับลิงก์ฟาร์มด้วยวิธีอนุมานไวยากรณ์กราฟ จะนำเสนอโดยเปรียบเทียบกับอัลกอริทึมการตรวจจับลิงก์ฟาร์มทั้ง 3 คือวิธีแอนตี้ทรัสทรานดักทิฟ และไวยากรณ์กราฟ การทดลองจะทำการแบ่งข้อมูลออกเป็น ชุดข้อมูลสอน และชุดข้อมูลทดสอบ ตามสัดส่วน 10% 20% 30% 40% และ 50% ของชุดข้อมูลสอน โดยใช้วิธีการ 10 ไขว้และหาค่าเฉลี่ยทั้งหมด 10 รอบ เมื่อกำหนดค่าเรียกคืนเป็น 50% และนำมาสร้างกราฟระหว่างสัดส่วนข้อมูลสอนกับค่าความแม่นยำ ได้ดังรูปที่ 4.16



รูปที่ 4.16 ค่าความแม่นยำกับสัดส่วนข้อมูลสอนเทียบกับอัลกอริทึมอื่น

จากนั้นเปรียบเทียบประสิทธิภาพของอัลกอริทึมระหว่างค่าความแม่นยำและค่าเรียกคืนในระดับต่างๆ กัน ตั้งแต่ 50% ถึง 100% โดยเก็บค่าเฉลี่ย 10 รอบแบบไขว้ และกำหนดอัตราส่วนชุดข้อมูลสอนเป็น 50% นำมาเขียนกราฟความสัมพันธ์ได้ดังรูปที่ 4.17



รูปที่ 4.17 ค่าความแม่นยำกับค่าเรียกคืนเทียบกับอัลกอริทึมอื่น

4.6 ผลการเปรียบเทียบประสิทธิภาพการตรวจจับลิงก์ฟาร์มที่มีคะแนนเพจแรงค์สูง

การเปรียบเทียบประสิทธิภาพการตรวจจับลิงก์ฟาร์มในกลุ่มที่มีคะแนนเพจแรงค์สูง ในถังที่ 1 ถึง 5 โดยจะเปรียบเทียบกับผลการทดลองในงานวิจัยตรวจจับลิงก์ฟาร์มโดยไวยากรณ์กราฟจากแบบจำลองลิงก์ฟาร์มที่เหมาะสม [16] เมื่อกำหนดสัดส่วนข้อมูลเป็น 50% สุ่มจากข้อมูลที่ติดฉลาก เปรียบเทียบค่าความแม่นยำและค่าเรียกคืนในแต่ละถังข้อมูล ณ จุดค่าความแม่นยำที่มากกว่าหรือเท่ากับค่าความแม่นยำที่กำหนดไว้ในงานวิจัยที่ตรวจจับโดยไวยากรณ์กราฟจากแบบจำลองลิงก์ฟาร์มที่เหมาะสม ดังแสดงในตารางที่ 4.5

ตารางที่ 4.5 เปรียบเทียบประสิทธิภาพในการตรวจจับลิงก์ฟาร์มที่มีคะแนนเพจแรงค์สูง

ถังที่	ค่าความแม่นยำ	ค่าเรียกคืน	
		ไวยากรณ์กราฟจากแบบจำลองลิงก์ฟาร์มที่เหมาะสม	ไวยากรณ์กราฟจากการอนุมานไวยากรณ์กราฟ
1	90.9090	0.7429	9.0000
2	96.2963	1.9316	25.4286
3	86.0465	5.4977	6.8571
4	84.4155	4.8291	84.0000
5	80.8695	6.9093	63.5000

4.7 วิเคราะห์ผลการทดลอง

ผลการทดลองประสิทธิภาพการตรวจจับลิงก์ฟาร์มบนโฮสต์กราฟในข้อมูลชุดข้อมูลทดสอบโฮสต์สแปม 5 ถังแรก พบว่าค่าความแม่นยำการตรวจจับสูงขึ้นตามสัดส่วนข้อมูลสอนและสูงสุดที่ 81.1308% ด้วยสัดส่วนข้อมูลสอน 50% ซึ่งค่าเรียกคืนเท่ากับ 79.18094% สำหรับค่าเรียกคืนสูงสุดเท่ากับ 80.80794 ซึ่งได้จากการสอนด้วยชุดข้อมูลสอนเพียง 30%

ผลการทดลองเมื่อเปรียบเทียบประสิทธิภาพการตรวจจับในค่าเรียกคืน 50% ด้วยสัดส่วนชุดข้อมูลสอนทั้ง 5 ระดับ คือ 10% 20% 30% 40% และ 50% พบว่าในช่วงแรกที่มีสัดส่วนชุดข้อมูลสอน 10% อัลกอริทึมการตรวจจับลิงก์ฟาร์มด้วยการอนุมานไวยากรณ์กราฟสามารถให้ค่าความแม่นยำที่สูงเทียบเท่ากับอัลกอริทึมการตรวจจับลิงก์ฟาร์มด้วยไวยากรณ์กราฟ และเมื่อเพิ่มสัดส่วนข้อมูลสอนที่มากขึ้นพบว่าอัลกอริทึมที่นำเสนอจะให้ค่าความแม่นยำที่สูงขึ้นเล็กน้อย ซึ่งแตกต่างจากอัลกอริทึมอื่นๆ ที่ให้ค่าความแม่นยำที่แตกต่างมากเมื่อเพิ่มสัดส่วนชุดข้อมูลสอน ดังนั้นอัลกอริทึมการตรวจจับลิงก์ฟาร์มด้วยการอนุมานไวยากรณ์กราฟที่นำเสนอนี้ หากใช้ข้อมูลสอนเพียงเล็กน้อยก็สามารถใช้ตรวจจับได้ดีเทียบเท่ากับการใช้สัดส่วนข้อมูลมาก

สำหรับผลการทดลองเมื่อเปรียบเทียบค่าความแม่นยำกับช่วงค่าความเรียกคืนตั้งแต่ 50% ถึง 100% พบว่า อัลกอริทึมที่นำเสนอมีค่าความแม่นยำต่ำกว่าอัลกอริทึมอื่นในช่วง 50% ถึง 65% แต่สามารถตรวจจับลิงก์ฟาร์มด้วยค่าความแม่นยำที่สูงกว่าทุกอัลกอริทึมที่ช่วงค่าเรียกคืน 65% ถึง 80% และช่วงที่ค่าเรียกคืนมากกว่า 80% อัลกอริทึมที่นำเสนอนี้ มีค่าความแม่นยำต่ำกว่าอัลกอริทึมการตรวจจับด้วยไวยากรณ์กราฟเล็กน้อย แต่อย่างไรก็ตาม เมื่อช่วงค่าเรียกคืนถึง 100% พบว่าค่าความแม่นยำของอัลกอริทึมที่นำเสนอสูงกว่าค่าความแม่นยำของอัลกอริทึมอื่น

ในส่วนการทดลองเปรียบเทียบประสิทธิภาพการตรวจจับลิงก์ฟาร์มในถังที่มีเพจแรงค์คะแนนสูง 5 ถังแรก พบว่าเมื่อกำหนดให้อัลกอริทึมทั้งสองตรวจจับด้วยค่าความแม่นยำที่มากกว่าหรือเท่ากัน และเปรียบเทียบค่าเรียกคืนของแต่ละถังข้อมูล พบว่าอัลกอริทึมการตรวจจับลิงก์ฟาร์มด้วยการอนุมานไวยากรณ์กราฟที่นำเสนอ สามารถตรวจจับโฮสต์สแปมได้มากกว่าในทุกถังข้อมูล

บทที่ 5
สรุปผลการวิจัย และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้พัฒนาการตรวจจับโครงสร้างลิงก์ฟาร์มโดยวิธีการอนุมานไวยากรณ์กราฟ ซึ่งลิงก์ฟาร์มที่ถูกสร้างขึ้นอย่างจงใจเพื่อเพิ่มคะแนนการจัดอันดับนี้ มักจะมีโครงสร้างลักษณะเฉพาะที่แตกต่างออกไปจากโครงสร้างเว็บปกติทั่วไป ซึ่งการวิเคราะห์ลักษณะเฉพาะในงานวิจัยนี้ใช้คะแนนเพจแรงค์และคะแนนทรังเคทเพจแรงค์ในการสร้างป้ายกำกับเพื่อติดลงบนเส้นเชื่อมของโฮสต์กราฟ ทั้งนี้เพื่อวิเคราะห์การถ่ายเทคะแนน และหาลักษณะเฉพาะของลิงก์ฟาร์ม จากผลการทดลองในการหาลักษณะของลิงก์ฟาร์มใน 2 ระดับระยะความห่างจากเว็บเป้าหมายพบว่า ลิงก์ฟาร์มจะมีโครงสร้างอินลิงก์และเอาท์ลิงก์ที่แตกต่างออกไปจากโครงสร้างเว็บปกติทั่วไป ดังนั้นจึงนำลักษณะดังกล่าวมาใช้เป็นไวยากรณ์กราฟในการตรวจจับลิงก์ฟาร์ม โดยการทดลองจะแบ่งข้อมูลออกเป็น 10 ถัง ตามคะแนนเพจแรงค์ และจะตรวจจับลิงก์ฟาร์มใน 5 ถังแรกเท่านั้น ทั้งนี้ลักษณะเฉพาะของโฮสต์สแปมที่มีคะแนนสูงและคะแนนต่ำมีความแตกต่างกัน อีกทั้งการกำจัดลิงก์ฟาร์มหรือเว็บสแปมที่มีคะแนนเพจแรงค์ ซึ่งปรากฏอยู่ในหน้าแรกๆ หรืออันดับต้นๆ ของผลการค้นคืน เป็นเป้าหมายที่สำคัญของการตรวจจับเว็บสแปม เพื่อให้โปรแกรมค้นหามีประสิทธิภาพยิ่งขึ้น

จากการทดลองประสิทธิภาพการเรียนรู้ด้วยสัดส่วนข้อมูลสอนต่างๆ พบว่าเมื่อใช้สัดส่วนข้อมูลสอนต่ำอัลกอริทึมการตรวจจับลิงก์ฟาร์มด้วยการอนุมานไวยากรณ์กราฟสามารถตรวจจับลิงก์ฟาร์มได้ด้วยค่าความแม่นยำเทียบเท่ากับอัลกอริทึมการตรวจจับลิงก์ฟาร์มด้วยไวยากรณ์กราฟ และเมื่อเพิ่มสัดส่วนการสอนเพิ่มขึ้นพบว่าค่าความแม่นยำการตรวจจับของอัลกอริทึมที่เสนอสูงขึ้นเพียงเล็กน้อย แสดงให้เห็นว่าอัลกอริทึมที่นำเสนอนี้สามารถใช้สัดส่วนข้อมูลสอนเพียงเล็กน้อยก็สามารถตรวจจับลิงก์ฟาร์มได้เกือบเทียบเท่าการใช้สัดส่วนข้อมูลสอนมาก ซึ่งแตกต่างจากอัลกอริทึมอื่นที่ต้องใช้ข้อมูลสอนด้วยสัดส่วนที่มาก เพื่อให้การตรวจจับมีประสิทธิภาพยิ่งขึ้น สำหรับการเปรียบเทียบประสิทธิภาพค่าความแม่นยำกับค่าเรียกคืนตั้งแต่ 50% ถึง 100% ด้วยอัตราส่วนชุดข้อมูลสอน 50% แม้ช่วงแรกค่าความแม่นยำจะต่ำกว่าอัลกอริทึมอื่น แต่อย่างไรก็ตามอัลกอริทึมที่นำเสนอมีแนวโน้มการลดลงของค่าความแม่นยำด้วยอัตราที่ต่ำกว่าอัลกอริทึมอื่น ทำให้ช่วงค่าเรียกคืนที่ 65% ถึง 80% อัลกอริทึมที่นำเสนอสามารถตรวจจับลิงก์ฟาร์มได้ด้วยค่าความแม่นยำที่สูงกว่าอัลกอริทึมอื่น

แม้ว่าไวยากรณ์กราฟจากการอนุมานที่นำเสนอในงานวิจัยนี้ ประสิทธิภาพการตรวจจับในช่วงค่าเรียกคืนตั้งแต่ 50% ถึง 100% มีบางช่วงที่มีค่าความแม่นยำที่ต่ำกว่าอัลกอริทึมการตรวจจับไวยากรณ์กราฟที่สร้างจากแบบจำลองลิงก์ฟาร์มที่เหมาะสม [16] แต่อย่างไรก็ตามจากตารางเปรียบเทียบประสิทธิภาพในการตรวจจับลิงก์ฟาร์มที่มีคะแนนเพจแรงค์สูง เมื่อพิจารณาค่าเรียกคืนในช่วงต้นที่มีค่าความแม่นยำสูงของการตรวจจับพบว่า อัลกอริทึมที่นำเสนอสามารถตรวจจับโฮสต์สแปมได้ในสัดส่วนที่มากกว่าทุกถังข้อมูล แสดงให้เห็นว่าไวยากรณ์กราฟจากการอนุมานที่นำเสนอมีประสิทธิภาพดีในการคัดกรองโฮสต์สแปมในช่วงแรกของการค้นคืน

5.2 **ข้อเสนอแนะ**

การตรวจจับลิงก์ฟาร์มในงานวิจัยนี้ ใช้การสร้างไวยากรณ์กราฟโดยการอนุมานไวยากรณ์กราฟ โดยพิจารณาจากสัดส่วนของรูปแบบเส้นเชื่อมรอบๆ โฮสต์เป้าหมาย จากโฮสต์เพื่อนบ้านที่แตกต่างกันเท่านั้น ยังไม่ได้พิจารณาปริมาณของเส้นเชื่อมที่เกิดจากโฮสต์เพื่อนบ้าน ซึ่งหากนำปริมาณเส้นเชื่อมมาพิจารณาด้วยแล้ว อาจทำให้ประสิทธิภาพในการตรวจจับลิงก์ฟาร์มหรือเว็บสแปมดียิ่งขึ้น

รายการอ้างอิง

[1] Gyongyi, Z., and Garcia-Nolina, H. Web Spam Taxonomy. Proceedings of the International Workshop on Adversarial Information Retrieval on the Web, 2005.

[2] Ntoulas, A., Najork, M., Manasse, M., and Fetterly, D. Detecting spam web pages through content analysis. Proceedings of the International conference on World Wide Web, pp.83-92. ACM Press, 2006.

[3] Becchetti, L., Castillo, C., Donato, D., Leonardi, S., and Baeza-Yates, R. Link-based characterization and detection of web spam, Proceedings of the International Workshop on Adversarial Information Retrieval on the Web, 2006.

[4] Becchetti, L., Castillo, C., Donato, D., Baeza-Yates, R., and Leonardi, S. Link analysis for web spam detection. ACM Transactions on the Web 2,1 (February 2008): 2:1-2:42.

[5] Abernethy, J., Chapelle, O., and Castillo, C. Web spam identification through content and hyperlinks, Proceedings of the International Workshop on Adversarial Information Retrieval on the Web, pp.41-44. 2008.

[6] Kiattikun Chobtham, Athasit Surarerks, and Arnon Rungsawang. Formalization of link farm structure using graph grammar. Proceedings of the IEEE International Conference on Advanced Information Networking and Applications, pp.904-910. 2008.

[7] Jeltsch, E., and Holder, L. Grammatical inference based on hyperedge replacement. Lecture Note in Computer Science 532 (1990): 461-474.

[8] Oates, T., Doshi, S., and Huang, F. Estimating maximum likelihood parameters for stochastic context-free graph grammars. Lecture Notes in Computer Science, 2835 (2003): 281-298.

[9] Kukluk, J.P., Holder, L.B., and Cook, D.J. Inference of Node Replacement Recursive Graph Grammars. Proceedings of the SIAM International Conference on Data Mining, pp.544-548. 2006.

[10] Kukluk, J.P., Holder, L.B., and Cook, D.J. Inference of node replacement graph grammar. Intelligent Data Analysis 11,4 (2007): 377-400.

[11] Castillo, C., Donato, D., Becchetti, L., and Boldi, P. A Reference collection for web spam. ACM SIGIR Forum 40,2 (2006): 11-24.

[12] Jonyer, I., Holder, L.B., and Cook, D.J. Concept formation using graph grammars. Proceedings of the KDD Workshop on Multi-Relational Data Mining, 2002.

[13] Pages, L., Brin, S., Motwani, R., and Winograd, T. The PageRank citation ranking: bring order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[14] Becchetti, L., Castillo, C., Donato, D., Leonardi, S., and Baeza-Yates, R. Using rank propagation and probabilistic counting for link-based spam detection. Proceedings of the Workshop on Web Mining and Web Usage Analysis, 2006.

[T15] Castillo, C., Donato, D. and Gionis, A. Know your neighbors: Web spam detection using the web topology. Proceedings of Annual international ACM SIGIR conference on Research and development in information retrieval, pp.423–430. 2007.

[16] เกียรติคุณ ชอบธรรม. การตรวจจับลิงก์ฟาร์มโดยใช้ไวยากรณ์กราฟ. วิทยานิพนธ์ปริญญามหาบัณฑิต, ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2551.

[17] Krishnan, V., and Raj, R. Web spam detection with anti-trust rank. Proceedings of the International Workshop on Adversarial Information Retrieval on the Web, pp.27-30. 2006.

[18] Zhou, D., Burges, C., and Tao, T. Transductive link spam detection. Proceedings of the International workshop on Adversarial Information Retrieval on the Web, pp.21-28. 2007.

[19] Jacek Kukluk. Inference of node and edge replacement graph grammar. Doctoral dissertation, Faculty of the Graduate School The University of Texas at Arlington, 2007.

[20] Ates, K., Kukluk, J., Holder, L., Cook, D., and Zhang, K. Graph grammar induction on structural data for visual programming. Proceedings of the IEEE International Conference on Tools with Artificial Intelligence, pp.232-242. 2006.

[21] Gyongyi, Z., Garcia-Molina, H., and Pedersen J. Combating web spam with TrustRank. Proceedings of the international conference on Very large data bases, pp.576–587. 2007.

[22] Benczur, A., Csalogany K., Sarlos, T., and Uher, M. SpamRank-fully automatic link spam detection. Proceedings of the International Workshop on Adversarial Information Retrieval on the Web, 2005.

ประวัติผู้เขียนวิทยานิพนธ์

นายวุฒิชัย วงศ์สารสิน เกิดเมื่อวันที่ 31 มีนาคม พ.ศ.2528 จบการศึกษาระดับมัธยมตอนต้นและตอนปลายจากโรงเรียนอุตรดิตถ์ จังหวัดอุตรดิตถ์ ต่อมาเข้าศึกษาต่อในระดับปริญญาบัณฑิต สาขาคณิตศาสตร์ประยุกต์ ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ จนสำเร็จการศึกษาในปีการศึกษา 2549 และเข้าศึกษาต่อในระดับปริญญามหาบัณฑิต สาขาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2550