ขั้นตอนวิธีการสำหรับการให้ค่าเรตติ้งและการวิเคราะห์เว็บไซต์อนาจาร

เกรียงกมล คำมา[1]* และ จักรกฤษณ์ เสน่ห์ นมะหุต[2]

บทคัดย่อ

ปัจจุบันอินเตอร์เน็ตมีบทบาทสำคัญในการอำนวยความสะดวกต่างๆ ไม่ว่าจะเป็นทางด้านการศึกษา การค้นคว้าหาความรู้ต่างๆ การทำธุรกรรมพาณิชอิเล็กทรอนิกส์เป็นต้น ซึ่งการใช้อินเตอร์เน็ตนั้นมีทั้งข้อดีและข้อเสีย โดยเฉพาะการเข้าถึงเว็บไซต์ที่ไม่เหมาะสมหรือเว็บไซต์อนาจารสำหรับเยาวชนนั้น ควรมีการเฝ้าระวังหรือป้องกันเพื่อไม่ให้ก่อเกิดปัญหาทางสังคมโดยเฉพาะปัญหาการกระทำชำเราทางเพศ แต่ด้วยจำนวนเว็บไซต์อนาจารที่เพิ่มอย่างทวีคูณในปัจจุบันทำให้เป็นเรื่องยากและซับซ้อนที่จะทำการกรองข้อมูลหรือวิเคราะห์ความเป็นอนาจารของเว็บไซต์ เนื่องจากตัวเว็บไซต์อนาจารเมื่อถูกตรวจพบจะมีการเปลี่ยนแปลงรูปแบบไป ในงานวิจัยนี้จึงได้พัฒนาวิธีการวิเคราะห์เว็บไซต์อนาจารด้วยเทคนิคที่ไม่ซับซ้อน ด้วยข่ายงานความหมาย (Semantic Network) คำอนาจารและการให้ค่าระดับความอนาจาร (Rating) ของคำอนาจารนั้นๆ เทียบกับคำที่พบในเว็บไซต์อนาจาร และวิเคราะห์หาค่าความเป็นอนาจารของเว็บไซต์ โดยใช้การวิเคราะห์จากโครงสร้างของเว็บไซต์ HTML Tags จากนั้นจะใช้ค่า Rating ของ Semantic Network คำอนาจารที่ได้สร้างไว้แล้วมาทำการคำนวณในแต่ละส่วนของโครงสร้างเพื่อนำมาวิเคราะห์ค่าความเป็นอนาจารของเว็บไซต์ จากการทดสอบระบบด้วยหลักการ F-measure กับเว็บไซต์ 150 เว็บไซต์ ประกอบไปด้วยเว็บไซต์ที่มีความเป็นอนาจาร เว็บไซต์ที่สื่อถึงความเป็นอนาจารแต่ไม่ได้เป็นเว็บไซต์อนาจาร และเว็บไซต์ที่ไม่มีความเป็นอนาจารเลย ตัวอย่างละ 50 เว็บไซต์ พบว่าประสิทธิภาพในการวิเคราะห์มีความถูกต้องสูงถึง 97 เปอร์เซ็นต์

คำสำคัญ: การจำแนกประเภทเว็บไซต์ การกรองเว็บไซต์ การวิเคราะห์เว็บไซต์ ข่ายงานความหมาย การสนับสนุนการตัดสินใจ

[1] นักศึกษา ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร

[2] ผู้ช่วยศาสตราจารย์ ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 0-5596-3223 อีเมล: khumma108@gmail.com

รับเมื่อ 19 พฤศจิกายน 2555 ตอบรับเมื่อ 21 กุมภาพันธ์ 2556

A Rating Algorithm for the Analysis of Pornographic Websites

Kriangkamon Khumma[1]* and Chakkrit Snae Namahoot[2]

Abstract

The internet plays an important role in such areas as education, knowledge distribution, and e-commerce services, which has both advantages and disadvantages. In particular, access to pornographic sites or sites that are not appropriate for children should be monitored. With the ever increasing number of pornographic Web sites, analyzing as well as filtering and blocking them is a complex and difficult task. Various inappropriate sites have been found to hide their contents, e.g. by advertising of other sites that are not pornographic or using words or jargon expression that do not appear to be pornographic. In this research, we have developed a simple method that can analyze pornographic websites by using a semantic list (or database) of inappropriate words for the rating of a Web site containing pornographic content. For the analysis of an inappropriate Web site we use the structural analysis of HTML tags. After that, the rating of the inappropriate words is used to perform the computation in each part of the site structure for pornography identification. The system tests were performed on 150 Web sites, from which 50 sites contain pornographic material, 50 sites contain indecent but not obscene material, and 50 sites contain appropriate material. It was found that the system can rate 97 percent of the Web sites correctly.

Keywords: Web Category, Web Filtering, Web Classification, Semantic Network, Decision Support

[1] Student, Department of Computer Science and Information Technology, Faculty of Science, Naresuan University.

[2] Assistant Professor, Department of Computer Science and Information Technology, Faculty of Science, Naresuan University.

* Corresponding Author, Tel. 0-5596-3223, E-mail: khumma108@gmail.com

Received 19 November 2012; Accepted 21 February 2013

1. บทนำ

การใช้งานอินเตอร์เน็ตในสังคมไทยกลายมาเป็นส่วนหนึ่งในชีวิตประจำวัน ใช้เพื่อการการศึกษา ความบันเทิง การทำธุรกรรมต่างๆ ซึ่งสามารถเข้าใช้ได้อย่างง่ายดายทั้งจากในหน่วยงานและจากที่พักอาศัย ซึ่งการเข้าใช้งานเว็บไซต์ต่างๆ ถ้าปราศจากการดูแลอาจจะทำให้การเข้าใช้งานเว็บไซต์บางอย่างไม่ก่อให้เกิดประโยชน์กับองค์กรและรบกวนการทำงานภายในองค์กรได้ สำหรับที่พักอาศัยที่มีเยาวชนอยู่อาจนำเข้าไปสู่เว็บไซต์ที่ไม่เหมาะสมกับช่วงอายุของเยาวชน เช่น เว็บไซต์ที่มีความรุนแรง เว็บไซต์ลามกอนาจาร ยิ่งไปกว่านั้น จากการสำรวจการใช้งานอินเตอร์แสดงถึงโอกาสที่จะเข้าสู่เว็บไซต์อนาจารนั้นมีโอกาสที่สูงมากทั้งแบบตั้งใจและไม่ตั้งใจ [1] เช่น ผลการสำรวจในประเทศสหรัฐอเมริกา ที่แสดงการเข้าใช้งานเว็บไซต์อนาจารของเยาวชนที่สูง [2] ส่วนในประเทศไทย ผลการสำรวจอิทธิพลของสื่อลามกในศูนย์ฝึกและอบรมเด็กและเยาวชนในเขตกรุงเทพและปริมณฑล พบว่าการกระทำความผิดทางเพศนั้นสาเหตุส่วนหนึ่งมาจากการเข้าถึงสื่อลามกที่เข้าถึงได้ง่ายตามความเจริญทางด้านเทคโนโลยี [3]

จากปัญหาเว็บไซต์อนาจาร ได้มีความพยายามในการกรองเว็บไซต์ที่ไม่เหมาะสม Caulkins และคณะ [4] อธิบายถึงสามวิธีหลักที่นิยมใช้คือ วิธีแรกคือการสร้างรายชื่อเว็บไซต์ที่สามารถเข้าใช้งานได้ (White List) และรายชื่อเว็บไซต์ที่ไม่อนุญาต (Black List) ไม่สามารถตอบสนองต่อเพิ่มจำนวนเว็บไซต์อย่างรวดเร็วได้ วิธีที่สองเป็นการระบุคำสำคัญแบบง่าย หากเว็บเพจใดมีคำที่ระบุจะไม่อนุญาตให้เข้าใช้งาน วิธีนี้มีปัญหาเรื่องการปิดกั้นมากเกินไป เช่น คำว่า “เต้านม” ทำให้ “มะเร็งเต้านม” ถูกปิดกั้นไปด้วย [5] สำหรับวิธีที่สามเป็นการใช้ค่า Rating ที่เป็นการบ่งบอกระดับความเหมาะสมในการเข้าใช้งานโดยกำหนดให้มีองค์กรที่คอยจัดการค่า Rating สำหรับแต่ละเว็บไซต์ ปัญหาที่ตามมาคือยังมีเว็บไซต์จำนวนมากที่ยังไม่มีการจัดอันดับค่า Rating เนื่องจากไม่ต้องการให้เว็บไซต์ของตนถูกคัดกรอง ต่อมาภายหลังใช้เทคนิคการจำแนกประเภทข้อความ (Text Classification, TC) เพื่อจำแนกประเภทเว็บไซต์ [5] เทคนิคที่นิยมใช้งาน ประกอบด้วย เทคนิคแรก Naïve Bayesian (NB) ที่ใช้ความน่าจะเป็นร่วมกันของเอกสารและประเภทเอกสารที่มีอยู่ เพื่อประมาณค่าความน่าจะเป็นว่าเอกสารควรจัดอยู่ประเภทใด เทคนิคที่สองคือ *k*-nearest neighbor (*k*-nn) จะพิจารณาค่า *k*-nn ของเอกสารแล้วประเมินว่าน่าจะเป็นเอกสารประเภทใดโดยจะใช้ค่าเกณฑ์ในการจำแนก Kwon และ Lee [6] นำ *k*-nn มาใช้วิเคราะห์เว็บไซต์โดยทำการวิเคราะห์เว็บเพจข้างเคียงของเว็บไซต์ แทนที่จะใช้เฉพาะเว็บเพจเริ่มต้นของเว็บไซต์เอง เพื่อดูว่าเว็บไซต์นั้นเป็นเว็บไซต์ประเภทใด เทคนิคถัดมาคือ Neural Network (NNet) เป็นโมเดลที่ออกแบบตามลักษณะการทำงานของระบบประสาทมนุษย์ NNet จะเรียนรู้รูปแบบโดยการปรับค่าน้ำหนักของแต่ละโหนดด้วยการเรียนรู้จากตัวอย่างที่ใช้สอนระบบใช้ในการทำ TC เพื่อคาดเดาประเภทของเอกสาร Lee และ Hui [7] นำเอาเทคนิค NNet มาคัดกรองเว็บไซต์อนาจาร โดยทำฝึกฝนระบบกับชุดตัวอย่างที่เตรียมไว้จากโครงสร้างบางส่วนประกอบด้วย Title, Display Content, Key Word เพื่อดำเนินการวิเคราะห์ เทคนิคต่อมาคือ Support Vector Machine (SVM) เป็นวิธีดำเนินงานโดยนำข้อมูลที่ใช้ในการวิเคราะห์แปลงให้อยู่ในรูปแบบข้อมูลเวกเตอร์ (Vector) จากนั้นจึงหาพื้นที่ระนาบเกิน (Hyperplane) เพื่อแบ่งคลาส (Class) Joachims [8] เป็นงานวิจัยแรกที่ประยุกต์ใช้ SVM ในการแก้ไขปัญหาการจำแนกประเภทข้อความ

สำหรับงานวิจัยในด้านการวิเคราะห์เว็บไซต์อนาจารที่ใช้การจำแนกประเภทข้อความภาษาไทยนั้น Polpinij และคณะ [9] ใช้เทคนิค SVM และ NB วิเคราะห์เว็บไซต์อนาจารทั้งภาษาไทยและภาษาอังกฤษ โดยเว็บไซต์จะอยู่ในรูปแบบถุงของคำ (Bag of Words) โดยไม่ให้ความสำคัญสำหรับโครงสร้างของ HTML Tags ที่พบคำ สำหรับการจำแนกประเภทเว็บไซต์ด้วยการให้ความสำคัญโครงสร้างที่พบคำจะมีประสิทธิภาพที่สูงกว่าแบบที่ไม่ให้ความสำคัญ [10] จากการศึกษาการเชื่อมโยงเว็บเพจนั้น เว็บเพจที่ถูกเชื่อมโยงไปมีโอกาสที่เป็น

เว็บเพจชนิดเดียวกันกับเว็บเพจที่วางตัวเชื่อมโยงค่อนข้างสูง [11] การวางเชื่อมโยงนั้นมีความสัมพันธ์ร่วมกันระหว่างตัวเชื่อมโยงกับตัวข้อมูลที่ทำการเชื่อมโยงไป [12] แต่ความสัมพันธ์ดังกล่าวจะเป็นความสัมพันธ์แบบหยาบที่กลุ่มในการจำแนกมีน้อย [13] เหมาะกับการวิเคราะห์เว็บไซต์อนาจารเนื่องจากกลุ่มในการจำแนกมีเพียงกลุ่มเดียวคือกลุ่มเว็บไซต์อนาจาร ในการวิจัยนี้จะจำแนกประเภทเว็บไซต์ที่เป็นเว็บไซต์อนาจารโดยให้ความสำคัญกับ HTML Tags และการเชื่อมโยงสู่เว็บเพจข้างเคียงจากเว็บเพจเริ่มต้นของเว็บไซต์ วิเคราะห์กับวิธีที่สร้างขึ้นมาเพื่อระบุคำอนาจารที่ค้นพบ วิเคราะห์เฉพาะข้อความภาษาไทย ด้วยวิธีที่ไม่ซับซ้อนที่พัฒนาขึ้นมาใหม่ ร่วมกับการสร้าง Semantic Network คำอนาจาร และทำการวิเคราะห์เว็บไซต์อนาจารกับ Semantic Network คำอนาจารที่สร้างไว้ ผ่านการตัดคำโดย SWATH และประเมินว่าเว็บไซต์ใดเป็นเว็บไซต์อนาจารกับเกณฑ์ (Threshold) ที่ได้กำหนด

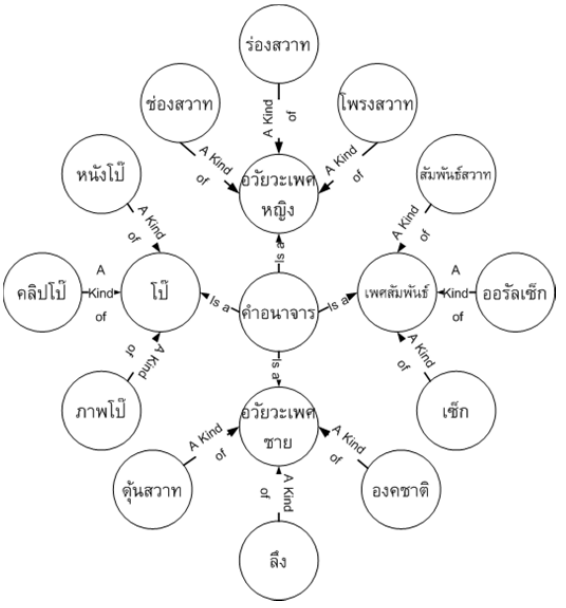
2. วิธีการวิจัย

ในการวิจัยนี้จะประกอบด้วยสามส่วน คือส่วนสร้าง Semantic Network คำอนาจาร ส่วนการตัดคำ และส่วนวิเคราะห์เว็บไซต์

2.1 การสร้าง Semantic Network คำอนาจาร

ประเทศไทยไม่มีองค์กรที่จัดระดับความรุนแรงทางภาษาไว้อย่างชัดเจน เช่น Caulkins และคณะ [4] ที่วิเคราะห์เว็บไซต์ภาษาอังกฤษ ประยุกต์ใช้การจัดกลุ่มคำจากองค์กรที่จัดระดับความรุนแรงของคำและแบ่งประเภทต่างๆ อย่างชัดเจน ฉะนั้นการศึกษาเพื่อวิเคราะห์เว็บไซต์อนาจารภาษาไทย จึงนำเอา Semantic Network ใช้สร้างกลุ่มคำอนาจารเพื่อนำไปใช้ในการวิเคราะห์เว็บไซต์

Semantic Network เป็นการบรรยายถึงการเชื่อมต่อระหว่างสถานีเชื่อมโยง (Node) ด้วยอาร์ก (Arc) โดยที่แต่ละ Node มีแนวคิดในการเชื่อมต่อกันโดยใช้ความสัมพันธ์แบบหลากหลาย [14] การวิจัยนี้ทำการ



รูปที่ 1 ตัวอย่างการสร้าง Semantic Network คำอนาจาร

สร้าง Semantic Network คำอนาจารด้วยความสัมพันธ์ระหว่างความหมายคำและ Rating เพื่อวัดค่าความใกล้เคียงของคำที่อยู่ใน Semantic Network เดียวกัน ดังรูปที่ 1 งานวิจัย [15] ตรวจสอบคำ (ชื่อ) ว่ามีความคล้ายคลึงกันหรือไม่โดยได้ตั้งค่าเกณฑ์ไว้ที่ 0.5 ถ้าค่านั้นมากกว่าหรือเท่ากับค่าเกณฑ์ 0.5 จะแสดงว่า ชื่อมีความคล้ายคลึงกัน แต่ถ้าหากน้อยกว่าแสดงว่าไม่มีความคล้าย ในการศึกษานี้ได้นำค่าเกณฑ์ดังกล่าวมาประยุกต์ใช้เพื่อวัดค่าความอนาจารของคำและวิเคราะห์ว่าเป็นเว็บไซต์อนาจาร สำหรับการสร้าง Semantic Network คำอนาจารในภาษาไทยยังต้องใช้ผู้เชี่ยวชาญในการกำหนดกำหนดความสัมพันธ์

ในรูปที่ 1 แสดงตัวอย่าง Semantic Network คำอนาจาร ได้แก่ โป๊ อวัยวะเพศชาย อวัยวะเพศหญิง และเพศสัมพันธ์ ซึ่งแต่ละคำจะมีคำที่มีความหมายเดียวกันส่วน Node ลำดับถัดไป เช่น คำว่าอวัยวะเพศชาย คำที่ความหมายเดียวกันได้แก่ ลึง ดุ้นสวาท องคชาติ

ขั้นตอนสร้าง Semantic Network คำอนาจารนั้นเริ่มจากรวบรวมรายการคำ สร้างเว็บไซต์เปิดให้ผู้ใช้

โพสต์คำที่คิดว่าเป็นคำอนาจาร จากนั้นจึงใช้ผู้เชี่ยวชาญทางภาษาตรวจสอบว่าคำใดเป็นคำอนาจารและคำนวณค่า Rating ด้วยการรวบรวมการใช้คำอนาจารจากเว็บไซต์อนาจารกลุ่มตัวอย่าง 50 เว็บไซต์ คำนวณตามสมการที่ (1)

$$R_i = \frac{Pw}{m} \quad (1)$$

R_i คือค่า Rating ของคำอนาจารของคำลำดับที่ i

Pw คือจำนวนเว็บไซต์อนาจารกลุ่มตัวอย่างที่พบการใช้คำอนาจารนั้น

m คือจำนวนเว็บไซต์อนาจารกลุ่มตัวอย่าง จำนวน 50 เว็บไซต์

ตารางที่ 1 ตัวอย่างการคำนวณค่า Rating คำอนาจาร

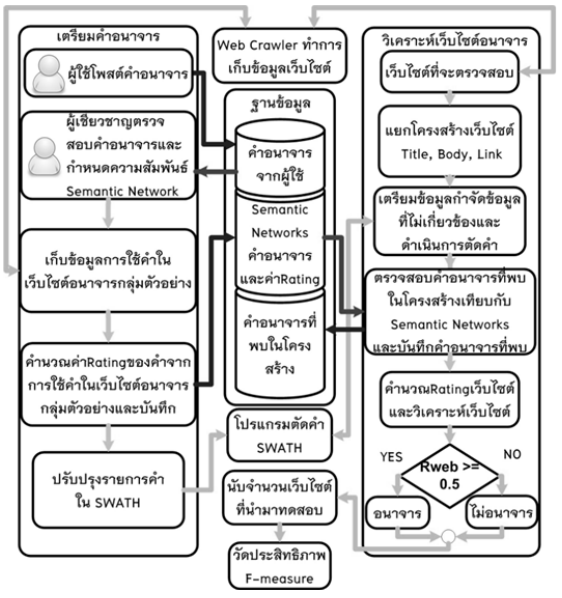
รายการคำ	Pw	R_i
โพรงสวาท	8	0.16
ช่องสวาท	28	0.56
ร่องสวาท	27	0.54
เซ็ก	39	0.78
ออรัลเซ็ก	39	0.78
สัมพันธ์สวาท	31	0.62
หนังโป๊	46	0.92
คลิปโป๊	45	0.9
ภาพโป๊	48	0.96
องคชาติ	20	0.4
ดุ้นสวาท	10	0.2
ลึง	22	0.44

ในตารางที่ 1 แสดงการคำนวณค่า Rating ของคำอนาจารจากรูปที่ 1 คำนวณกับเว็บไซต์อนาจารกลุ่มตัวอย่าง ตรวจการใช้งานคำว่าพบในกี่เว็บไซต์และคำนวณตามสมการที่ (1) เช่น คำว่า “โพรงสวาท” มีการใช้คำในเว็บไซต์ 8 เว็บไซต์ คือ Pw ค่า และ m คือ 50 เมื่อคำนวณแล้วจะได้ค่า R_i คือ 0.16 ซึ่งเป็นค่า Rating สำหรับ “โพรงสวาท” และทำการบันทึกข้อมูลคำ ความสัมพันธ์ และค่า Rating ลงในฐานข้อมูลระบบ MySQL เพื่อใช้ในการวิเคราะห์เว็บไซต์อนาจารต่อไป

2.2 การตัดคำ

รูปแบบของภาษาไทยจะเขียนข้อความติดกันในประโยค ไม่มีการแบ่งคำ เหมือนรูปแบบของภาษาอังกฤษที่ใช้ช่องว่างในการแบ่งคำ เพื่อแก้ปัญหานี้ จึงใช้โปรแกรม Smart Word Analysis for Thai (SWATH) ในการตัดคำภาษาไทยในระหว่างการวิเคราะห์ SWATH มีขั้นตอนวิธีสำหรับทำงานสามวิธี ได้แก่ วิธีแรก เป็นวิธีการตัดคำแบบยาวที่สุด (Longest Matching) เป็นการตัดคำโดยให้คำที่ยาวที่สุดที่ค้นพบได้จากพจนานุกรมของโปรแกรม เริ่มจากทางซ้ายสุดของประโยคก่อน เช่น “ฉันนั่งตากลมหน้าบ้าน” สามารถตัดคำได้สองแบบที่คือ “ฉัน|นั่ง|ตา|กลม|หน้า|บ้าน” และ “ฉัน|นั่ง|ตาก|ลม|หน้า|บ้าน” ถ้าใช้ Longest Matching จะได้แบบที่สองคือคำว่า “ตาก” เพราะเป็นคำที่ยาวที่สุดที่พบในพจนานุกรม วิธีที่สอง วิธีการตัดคำแบบสอดคล้องมากที่สุด (Maximal Matching) เป็นการตัดคำที่จะให้คำที่ได้จากการตัดคำมีจำนวนน้อยที่สุด เช่น “หามเหสี” สามารถตัดคำได้สองแบบ “หาม|เห|สี” และ “หา|มเหสี” ถ้าตัดคำโดย Maximal Matching จะได้แบบที่สองที่มีจำนวนคำที่น้อยกว่าคือจำนวน 2 คำ แต่ในกรณีสำหรับ “ฉันนั่งตากลมหน้าบ้าน” จะตัดได้จำนวน 6 คำเท่ากันทั้งสองแบบ แต่เมื่อใช้ Longest Matching ด้วยก็จะตัดคำเป็นคำว่า “ฉัน|นั่ง|ตาก|ลม|หน้า|บ้าน” เพราะ “ตาก” ยาวกว่าคำว่า “ตา” และวิธีที่สาม คือ Part of Speech Bigram ที่ใช้ลำดับหน้าที่ของคำและคำที่มี 2 คำติดกันในข้อความ (Bigram) สร้างจากคำโดยใช้วิธีการเลื่อนไปทีละสองคำ

ในการศึกษานี้ใช้ SWATH ตัดคำด้วยวิธี Longest Matching และ Maximal Matching



รูปที่ 2 แสดงการทำงานของระบบ

2.3 การวิเคราะห์เว็บไซต์

การวิเคราะห์เว็บไซต์จะทำงานในรูปแบบ Web Application โดยใช้ภาษา PHP ทำงานกับฐานข้อมูลเชิงสัมพันธ์ MySQL การทำงานดังรูปที่ 2

รูปที่ 2 แสดงการทำงานของระบบ ประกอบด้วยส่วนที่เตรียมคำอนาจาร ฐานข้อมูล โปรแกรม SWATH และส่วนวิเคราะห์เว็บไซต์

สำหรับส่วนวิเคราะห์จะรับข้อมูล URL ที่ใช้ในการวิเคราะห์แล้วดำเนินการเก็บข้อมูลที่จะใช้ในการวิเคราะห์ คือเว็บเพจเริ่มต้นของเว็บไซต์ เว็บเพจข้างเคียงที่เชื่อมโยงผ่าน Link ที่วางในเว็บเพจเริ่มต้น จากนั้นแยกโครงสร้างที่ใช้ในการวิเคราะห์ ดำเนินการ กำจัดข้อมูลที่ไม่เกี่ยวข้องในการวิเคราะห์ ทำการตัดคำ คำนวณค่า Rating ของเว็บไซต์อนาจาร และประเมินกับค่าเกณฑ์

ขั้นตอนวิธีในการวิเคราะห์เว็บไซต์อนาจารมีการทำงานดังนี้

1) ระบุ URL ของเว็บไซต์ที่ต้องการดำเนินการวิเคราะห์

2) ดำเนินการเก็บข้อมูลของ URL ที่ต้องการวิเคราะห์และดำเนินการบันทึกเป็นไฟล์ไว้ จะใช้เว็บเพจเริ่มต้นเมื่อเรียกด้วย URL ของเว็บไซต์

3) แยกข้อมูล HTML Tag ที่จะใช้ในการวิเคราะห์ Title, Body, Link มีวิธีการวิเคราะห์ ดังนี้

3.1) ดึงข้อมูลจาก Tag ที่ใช้ในการวิเคราะห์

3.2) เตรียมข้อมูล กำจัดข้อมูลที่ไม่เกี่ยวข้อง เช่น ภาษาอังกฤษ ตัวเลข จากนั้นดำเนินการตัดคำภาษาไทยโดยโปรแกรม SWATH

3.3) ตรวจสอบคำใดว่าเป็นคำอนาจารกับ Semantic Network คำอนาจารที่เตรียมไว้ และคำนวณค่า Rating บันทึกผลการคำนวณของแต่ละโครงสร้างลงในฐานข้อมูล สำหรับค่า Rating ของเว็บไซต์คำนวณตามสมการที่ (2)

$$R_{web} = \frac{t+b+l}{m} \tag{2}$$

$$t = \frac{\sum_{i=1}^{p}(tR_i)}{p} \tag{3}$$

$$b = \frac{\sum_{i=1}^{q}(bR_i)}{q} \tag{4}$$

$$l = \frac{\sum_{1}^{z}\left(\sum_{i=1}^{x}\sum_{j=1}^{y}\left(\frac{t\acute{R}_i}{x}+\frac{b\acute{R}_i}{y}\right)\right)}{z} \tag{5}$$

R_{web} คือค่า Rating ของเว็บไซต์

t คือค่า Rating ของส่วน Title ที่คำนวณจากสมการที่ (3)

b คือค่า Rating ของส่วน Body ที่คำนวณจากสมการที่ (4)

m คือจำนวนโครงสร้างที่ใช้ในการคำนวณคือ 3 โครงสร้าง

tR_i คือค่า Rating ของคำอนาจารลำดับที่ i แต่ละคำที่พบใน Title

p คือจำนวนคำอนาจารที่พบใน Title

bR_i คือค่า Rating ของคำอนาจารลำดับที่ i แต่ละคำที่พบใน Body

q คือจำนวนคำอนาจารที่พบใน Body

l คือค่า Rating เว็บเพจข้างเคียงทั้งหมดที่เชื่อม

โยงเว็บเพจเริ่มต้นของเว็บไซต์คำนวณได้จากสมการที่ (5)

$t\acute{R}_i$ คือค่า Rating ของคำอนาจารลำดับที่ i แต่ละคำที่พบใน Title ของเว็บเพจทุกเพ็จที่เชื่อมโยงจากเว็บเพจเริ่มต้น

x คือจำนวนคำอนาจารทั้งหมดในส่วนของ Title ของเว็บเพจข้างเคียงทุกเพ็จเชื่อมโยงจากเว็บเพจเริ่มต้น

$b\acute{R}_i$ คือค่า Rating คำอนาจารลำดับที่ i แต่ละคำที่พบใน Body ของเว็บเพจทุกเพ็จที่เชื่อมโยงจากเว็บเพจเริ่มต้น

y คือจำนวนคำอนาจารทั้งหมดในส่วนของ Body ของเว็บเพจทุกเพ็จข้างเคียงเชื่อมโยงจากเว็บเพจเริ่มต้น

z คือจำนวนเว็บเพจข้างเคียงทั้งหมดที่สามารถเชื่อมโยงจากเว็บเพจเริ่มต้นของเว็บไซต์

4) บันทึกค่า Rating ของเว็บไซต์ที่คำนวณได้และคำอนาจารทั้งหมดที่พบระหว่างการวิเคราะห์ จากนั้นประเมินว่าเป็นเว็บไซต์อนาจารหรือไม่ โดยใช้ Rating ของเว็บไซต์เปรียบเทียบกับค่าเกณฑ์ที่ได้กำหนดไว้และรายงานผลการวิเคราะห์ เปรียบเทียบกับเงื่อนไขดังนี้

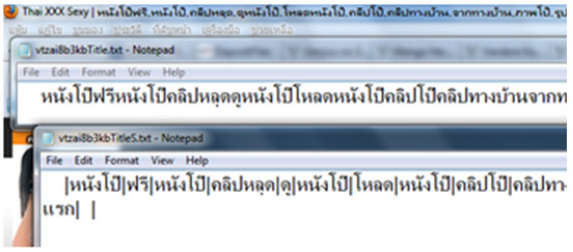
ค่า Rating ของเว็บไซต์ตั้งแต่ 0 ไม่ถึง 0.5 ผลการวิเคราะห์ “ไม่น่าจะเป็นเว็บไซต์อนาจาร” แต่ถ้าค่า Rating เว็บไซต์มีค่าตั้งแต่ 0.5 ผลการวิเคราะห์ “น่าจะเป็นเว็บไซต์อนาจาร”

3. ผลการทดสอบระบบและวิจารณ์

ในการทดสอบระบบจะแบ่งผลการทดสอบระบบออกเป็น 3 ส่วน คือผลการตัดคำโดยโปรแกรม SWATH ผลการทำงานของระบบวิเคราะห์เว็บไซต์ และผลการวัดประสิทธิภาพการทำงาน

3.1 ผลการตัดคำโดยโปรแกรม SWATH

โปรแกรม SWATH มีรายการคำภายในโปรแกรมเพื่อนำมาใช้กับการตัดคำ ทำให้มีปัญหาเรื่องการตัดคำที่เป็นอนาจาร และทำให้การวิเคราะห์นั้นมีประสิทธิภาพลดลง จึงต้องมีการปรับปรุงรายการคำ ซึ่งโปรแกรม SWATH อนุญาตให้ผู้ใช้สามารถปรับปรุงรายการคำ



รูปที่ 3 แสดงตัวอย่างการตัดคำในส่วน Title โดย SWATH ที่ปรับรายการคำแล้ว

ได้ จึงสามารถเพิ่มรายการคำใน Semantic Network คำอนาจาร เข้าไปในรายการคำของโปรแกรม

รูปที่ 3 แสดงตัวอย่างเตรียมข้อมูลในส่วนวิเคราะห์ดำเนินการตัดคำด้วยโปรแกรม SWATH การตัดคำโดย SWATH นั้นสามารถกำหนดอักขระที่ใช้ในการแบ่งคำสำหรับการทดสอบนี้ใช้ “|” ในการแบ่งคำ เมื่อได้คำแล้วจะเริ่มทำการวิเคราะห์เว็บไซต์อนาจารในขั้นตอนต่อไป

3.2 ผลการทดสอบวิเคราะห์เว็บไซต์

ระบบที่พัฒนานั้นสามารถวิเคราะห์เว็บไซต์ตามขั้นตอนวิธีที่ได้อธิบายในส่วนวิธีการวิจัย เมื่อระบบตัดคำแล้วจะทำการวิเคราะห์แต่ละโครงสร้าง ตัวอย่างเช่นในส่วน Title ของเว็บเพจเริ่มต้นของเว็บไซต์ที่ตัดคำในรูปที่ 3 คำที่เป็นคำอนาจารและค่า Rating ได้แก่ [หนังโป๊] =0.92 [หนังโป๊]=0.92 [คลิปหลุด]=0.75 [หนังโป๊]=0.92 [หนังโป๊]=0.92 [คลิปโป๊]=0.9 [ภาพโป๊]=0.9 [รูปโป๊]=0.9 จากสมการที่ (3) จำนวนคำอนาจาร 8 คำ p=8 และค่า $\sum_{i=1}^{p}(t\acute{R}_i)$ =7 จะได้ค่าของ t = 0.89 โครงสร้าง Body ใช้การคำนวณที่คล้ายกับ Title ดังสมการที่ (4) แต่ Link จะเชื่อมโยงไปเว็บเพจข้างเคียงทั้งหมดจาก Link ที่วางในเว็บเพจเริ่มต้นของเว็บไซต์ และคำนวณจากโครงสร้าง Title, Body ในเว็บเพจข้างเคียงทั้งหมด ตามสมการที่ (5) เมื่อได้ค่า Rating ของแต่ละโครงสร้างจะคำนวณด้วยสมการที่ (2) และประเมิน Rating เว็บไซต์ Rweb กับค่าเกณฑ์ ระบบรายงานผลการทำงานดังรูปที่ 4 จากรูปที่ 4 ตัวอย่างในการ

รูปที่ 4 แสดงตัวอย่างส่วนวิเคราะห์เว็บไซต์

วิเคราะห์เว็บไซต์ ระบบจะแสดงค่า Rating ที่คำนวณได้ดังนี้ Title 0.8975 Body 0.8416 Link 0.7653 และจำนวนเว็บเพจข้างเคียงที่เก็บข้อมูล ค่า Rating ของเว็บไซต์ คือ 0.8348 ผลการประเมินกับค่าเกณฑ์น่าจะเป็นเว็บไซต์อนาจาร

ดำเนินการทดสอบระบบ โดยทำการวิเคราะห์เว็บไซต์จำนวน 150 เว็บไซต์แบ่งเป็น 3 กลุ่ม ได้แก่ กลุ่มแรก คือเว็บไซต์อนาจารจำนวน 50 เว็บไซต์ รายชื่อเว็บไซต์อนาจารนี้ได้จากผู้ใช้ที่ร่วมโพสต์รายชื่อเว็บไซต์อนาจาร เว็บไซต์ทดสอบจะไม่ซ้ำกับเว็บไซต์อนาจารกลุ่มตัวอย่างที่ใช้คำนวณค่า Rating ของคำ กลุ่มที่สอง คือเว็บไซต์ที่ไม่ใช่เว็บไซต์อนาจารจำนวน 50 เว็บไซต์ กลุ่มที่สาม คือเว็บไซต์ที่มีคำอนาจารภายในเว็บไซต์แต่ไม่ได้เป็นเว็บไซต์อนาจาร 50 เว็บไซต์ เช่น เว็บไซต์ที่ให้ความรู้เรื่องสุขภาพทางเพศ ที่มีคำที่เกี่ยวข้องบางส่วนแต่เว็บไซต์เองไม่มีเป้าหมายเชิงอนาจาร ได้ผลการทดสอบดังตารางที่ 2

ในตารางที่ 2 ผลการทดสอบการวิเคราะห์กับเว็บไซต์แต่ละกลุ่ม ด้วยระบบที่พัฒนาขึ้นเพื่อวัดประสิทธิภาพการวิเคราะห์

ผลการวิเคราะห์เว็บไซต์จำนวน 150 เว็บไซต์สามกลุ่ม สำหรับกลุ่มเว็บไซต์อนาจารนั้น ระบบวิเคราะห์ผิดพลาดจำนวน 10 เว็บไซต์ แบ่งเป็น 2 กรณีหลัก คือ กรณีแรก คำอนาจารภาษาอังกฤษใช้แทนคำอนาจารภาษาไทยในบางโครงสร้างจึงไม่สามารถวิเคราะห์ได้ เพราะการศึกษานี้ขอบเขตคำอนาจารภาษาไทย เช่น เว็บไซต์ http://xstory.aoi2you.com ใช้คำอนาจารภาษาอังกฤษแทนคำอนาจารภาษาไทยใน Title ทำให้ t=0 ในส่วนของ Body วิเคราะห์ได้ตามปกติ b=0.81 แต่ในส่วนเว็บข้างเคียงทั้งหมดจะมีลักษณะที่คล้ายกันใช้คำอนาจารภาษาอังกฤษแทนคำอนาจารภาษาไทยจำนวนมากจึงได้ค่า l=0.63 ที่ไม่สูงมากทำให้ R_{web}=0.48 เมื่อวิเคราะห์กับเกณฑ์ที่ตั้งไว้จึงไม่ใช่เว็บไซต์อนาจาร กรณีสองวิเคราะห์ผิดพลาดเกิดจากส่วนของ Body ของเว็บไซต์จะมีคำอนาจารน้อยมากหรือไม่มีคำอนาจารเพราะจะใช้เป็นภาพนิ่งและภาพเคลื่อนไหวแทน ซึ่งเป็นภาพอนาจารอย่างชัดเจนทำให้ b ที่ต่ำมาก ถึงแม้ส่วนของ t ที่มีค่าสูง และเว็บเพจข้างเคียงที่เชื่อมโยงผ่าน Link ก็จะเป็นเช่นเดียวกันทำให้ l ที่ต่ำลงด้วย เช่น เว็บไซต์ http://www.guchob.com ที่ t สูงถึง 0.83 แต่ b นั้นต่ำมากเพียง 0.2 และยิ่ง l เพียง 0.31508 ยิ่งทำให้ค่า R_{web} มีค่าลดลงอีกเพียง 0.45 ทำให้ระบบวิเคราะห์ว่าไม่เป็นเว็บไซต์อนาจาร

ตารางที่ 2 ผลการทดสอบวิเคราะห์เว็บไซต์แต่ละกลุ่ม

ประเภทเว็บไซต์	จำนวนเว็บไซต์	ผลการวิเคราะห์		F-measure
		อนาจาร	ไม่อนาจาร	
เว็บไซต์อนาจาร	50	40	10	89%
เว็บไซต์ที่ไม่ใช่อนาจาร	50	0	50	100%
เว็บไซต์ที่มีคำอนาจารแต่ไม่ใช่อนาจาร	50	0	50	100%
รวม/ค่าเฉลี่ย	150	40	110	97%

3.3 การวัดประสิทธิภาพของระบบ

การวัดประสิทธิภาพของระบบด้วยค่า F-measure ที่ใช้วัดประสิทธิภาพของการคืนค้น [16]

$$F-measure = \frac{2*Precision*Recall}{Precision+Recall} * 100 \qquad (6)$$

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \quad (7)$$

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \quad (8)$$

สำหรับการวัดประสิทธิภาพด้วยค่า F-measure ของกลุ่มเว็บไซต์อนาจารนั้นค่า True Positive (TP) คือ 40 ค่า False Negative (FN) คือ 10 False Positive (FP) คือ 0 สำหรับกลุ่มที่ไม่ใช่เว็บไซต์อนาจารและกลุ่มที่มีคำอนาจารแต่ไม่ใช่เว็บไซต์อนาจารทั้งสองกลุ่มจะมีค่า TP คือ 50 FP และ FN คือ 0 และเมื่อคำนวณหาค่า F-measure ของแต่ละกลุ่มแล้วคำนวณค่า F-measure เฉลี่ย จะได้ค่าประสิทธิภาพของระบบประมาณ 97 เปอร์เซ็นต์

4. อภิปรายผลและสรุป

ในการศึกษานี้ได้เสนอวิธีการวิเคราะห์เว็บไซต์อนาจารโดยเทคนิคการให้ค่า Rating คำอนาจารและเว็บไซต์ คำอนาจารที่ใช้ในการวิเคราะห์ได้จากการรวบรวมคำอนาจารผ่านการตรวจสอบด้วยผู้เชี่ยวชาญและคำนวณค่า Rating ของคำกับกลุ่มตัวอย่างเว็บไซต์อนาจาร จากนั้นทำการวิเคราะห์เว็บไซต์ ขอบเขตวิเคราะห์คำภาษาไทยใน HTML Tags ประกอบด้วย Title, Body และเว็บเพจข้างเคียงที่เชื่อมโยงผ่าน Link ในเว็บเพจเริ่มต้นของเว็บไซต์ ด้วยเทคนิคที่สร้างขึ้นเพื่อตรวจสอบว่าเป็นเว็บไซต์อนาจาร ทดสอบกับเว็บไซต์ 3 กลุ่ม จากผลการทดสอบพบว่าได้ค่า F-measure ประมาณ 97 เปอร์เซ็นต์

สำหรับการศึกษาต่อไป จะประยุกต์ประมวลผลรูปภาพภาพ (Image Processing) เพื่อช่วยแก้ปัญหาที่เว็บไซต์อนาจารบางเว็บไซต์ ใช้ภาพสื่อความหมายด้านอนาจารอย่างชัดเจนแทนข้อความ และพัฒนาระบบให้สามารถวิเคราะห์คำอนาจารภาษาอังกฤษ และเพิ่มประสิทธิภาพการวิเคราะห์โดยใช้อาศัยคุณลักษณะของคำ (Featured Base) เข้าช่วยโดยการพิจารณาจากบริบท (Context) รอบข้างเพื่อให้ทราบหน้าที่ของคำอย่างชัดเจนมากยิ่งขึ้นและใช้ตรรกศาสตร์คลุมเครือ (Fuzzy Logic) เพื่อพิจารณาความหมายของประโยค ปรับปรุงการให้ค่า Rating ของคำ กรณีคำที่ไม่เจอใน Semantic Network คำอนาจารที่สร้างขึ้น แต่มีความถี่ในการใช้งานเว็บไซต์อนาจารสูง หรือใช้ปรับค่า Rating คำอนาจารที่มีในระบบจากความถี่การใช้คำดังกล่าว เพื่อให้มีประสิทธิภาพในการวิเคราะห์ที่สูงขึ้น

เอกสารอ้างอิง

[1] Survey, (2002). The Stat on internet pornography, OnlineMBA. [Online]. Available: http://www.onlinemba. com/ blog/ the-stats-on-internet-porn/

[2] Survey, (2002) National Public Radio, the Kaiser Family Foundation, and Harvard's Kennedy School of Government. [Online]. Available: http://www. npr.org/ programs/specials/poll/technology

[3] Yapaporn Piwapong, "The influence of pornography on male juvenile sexual abuse : a case study of juvenile vocational training center for boys in the Bangkok metropolitan and vicinity area," *Academic Services Journal,* Vol.22, No.3, pp.13-39, 2001 (in Thai)

[4] J.P. Caulkins, W. Ding, G. Duncan, R. Krishnan, E. Nyberg, "A method for managing access to web page: Filtering by Statiscal Classification (FSC) applied to text," *Decision Support System,* vol. 42, pp. 144-161, 2006.

[5] M. Chau and H. Chau, "A machine learning a pproach to web page filtering using content and structure analysis," *Decision Support System,* vol. 44, pp. 482-494, 2008.

[6] O. Kwon and J. Lee, "Text categorization based on k-nearest neighbor approach for Web site classification," *Information Processing & Management,* vol. 39 pp. 25-44, 2003.

[7] P.Y. Lee and S.C. Hui, "Neural network for web

content filtering," *IEEE Intell. Syst.* vol.17, pp. 48-57, Sep 2002.

[8] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," *Proc. of the European Conf. on Mach. Learning,* Berlin, pp. 137-142, 1998.

[9] J. Polpinij, A. Chotthanom, C. Sibunruang, R. Chamchong, S. Puangpronpitag, "Content-Based Text Classifiers for Pornographic Web Filtering." *IEEE Int. Conf. on Syst., Man, and Cybern.* Taipei, Taiwan, October 2006.

[10] K. Golub and A. Ardo, "Importance of HTML structural elements and metadata in automated subject classification," *Proc. of the 9th European Conf. on Research and Advanced Technology for Digital Libraries (ECDL).* Lecture Notes in Comput. Sci. vol. 3652. Springer, Berlin, Germany, pp. 363-378, 2005.

[11] S. Chakrabatri, *Mining the Web: Discovering Knowledge from Hypertext Data,* Morgan Kaufmann, San Francisco, CA, 2003.

[12] F. Menczer, "Mapping the semantics of Web text and links," *IEEE Internet Compt,* vol.9, pp. 27-36, 2004.

[13] X. Qi and B.D. Davison, "Web Page Classification: Features and Algorithms," *ACM Computing Surveys,* vol. 41(2), 2009.

[14] J. H. Lee, M. H. Kim, Y. J. Lee, "Information retrieval based on conceptual distance in IS-A hierarchies," *J. of Documentation,* vol.49 (2), pp. 188-207, 1993.

[15] C. Snae and B. Diaz, "An Interface for mining Genealogical Nominal Data Using the Concept of Linkage and a Hybrid Name Matching Algorithm," *J. of 3D-Forum Society,* vol.16, pp. 147-152, 2002.

[16] V. Vapnik, *Statistical Learning Theory,* New York: John Wesley & Sons, 1998.