การพัฒนาระบบคัดกรองเว็บไซต์ที่ไม่เหมาะสมในระดับมัธยมศึกษา

โดย

นายชาญพัฒณ์ ภินันท์รัชต์ธร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

ภาควิชาคอมพิวเตอร์

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ปีการศึกษา 2552

ลิขสิทธิ์ของบัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

การพัฒนาระบบคัดกรองเว็บไซต์ที่ไม่เหมาะสมในระดับมัธยมศึกษา

โดย
นายชาญพัฒณ์ ภินันท์รัชต์ธร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
ภาควิชาคอมพิวเตอร์
บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร
ปีการศึกษา 2552
ลิขสิทธิ์ของบัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

SYSTEM DEVELOPING OF WEB CONTENT FILTERING IN SECONDARY SCHOOL

By

Chanpat Pinunratchathon

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

Department of Computing

Graduate School

SILPAKORN UNIVERSITY

2009

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร อนุมัติให้วิทยานิพนธ์เรื่อง “การพัฒนาระบบคัดกรองเว็บไซต์ที่ไม่เหมาะสมในระดับมัธยมศึกษา” เสนอโดย นายชาญพัฒณ์ ภินันท์รัชต์ธร เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

……..

(รองศาสตราจารย์ ดร.ศิริชัย ชินะตังกูร)

คณบดีบัณฑิตวิทยาลัย

วันที่.....เดือน..... พ.ศ.....

ผู้ควบคุมวิทยานิพนธ์

ผู้ช่วยศาสตราจารย์ ดร.ปานใจ ธารทัศนวงศ์

คณะกรรมการตรวจสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร.จันทนา ผ่องเพ็ญศรี)

...../...../.....

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.กฤษณะ ชินสาร)

...../...../.....

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ปานใจ ธารทัศนวงศ์)

...../...../.....

48309322 : สาขาวิชาเทคโนโลยีสารสนเทศ
คำสำคัญ : การคัดกรองเว็บไซต์ที่ไม่เหมาะสม / ซัพพอร์ตเวกเตอร์แมชชีน / การวิเคราะห์องค์ประกอบหลัก

ชาญพัฒณ์ ภินันท์รัชต์ธร : การพัฒนาระบบคัดกรองเว็บไซต์ที่ไม่เหมาะสมในระดับมัธยมศึกษา. อาจารย์ที่ปรึกษาวิทยานิพนธ์ : ผศ.ดร.ปานใจ ธารทัศนวงศ์. 154 หน้า.

วิทยานิพนธ์นี้ได้พัฒนาขึ้นเพื่อการคัดกรองเว็บไซต์ที่ไม่เหมาะสมภายใต้ระบบปฏิบัติการลีนุกซ์ โดยใช้โปรแกรม Squid ซึ่งมี ACL (Access Control List) เป็นตัวควบคุมการใช้งานเว็บไซต์ และได้นำเอาอัลกอริทึมการวิเคราะห์องค์ประกอบหลัก PCA (Principal Component Analysis) และอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน SVM (Support Vector Machine) มาใช้ในการคัดกรองเว็บไซต์ เพื่อช่วยควบคุมการใช้งานอินเทอร์เน็ตภายในโรงเรียนให้นักเรียนได้รับข้อมูลที่เหมาะสม

หลักการสร้างระบบคัดกรองเว็บไซต์ที่ไม่เหมาะสมนั้น ได้ใช้โครงสร้างของเว็บไซต์ตาม TAG ต่าง ๆ ดังนี้ META, IMG, A HREF, SCRIPT, TITLE และ BODY มาเป็นองค์ประกอบในการสร้างระบบคัดกรอง โดยนำเว็บไซต์ไม่เหมาะสมมาหาความสัมพันธ์ และใช้ PCA มาพิจารณาค่าความแปรปรวนในแต่ละองค์ประกอบเพื่อสร้างตัวแบบ จากนั้นทำการทดสอบโดยนำเว็บไซต์ความรุนแรง ยาเสพติด 200 เว็บไซต์ และเว็บไซต์ลามกอนาจาร 200 เว็บไซต์ มาทดสอบกับระบบ ผลของการศึกษาตัวแบบพบว่าระบบที่สร้างขึ้นสามารถแบ่งแยกเว็บไซต์ไม่เหมาะสมออกจากเว็บไซต์ปกติได้ โดยมีความถูกต้องสำหรับกลุ่มที่มีความรุนแรง ยาเสพติด 89.5 % และกลุ่มลามกอนาจาร 94.5% และทำการเปรียบเทียบวิธีการคัดกรองเว็บไซต์ที่ไม่เหมาะสมกับอัลกอริทึม SVM โดยใช้ข้อมูลทดสอบชุดเดียวกัน พบว่าวิธี SVM มีความถูกต้องในการคัดกรองสำหรับกลุ่มเว็บไซต์ความรุนแรง ยาเสพติด 89 % และกลุ่มเว็บไซต์ลามกอนาจาร 91% ในขณะที่ วิธี PCA สามารถแบ่งกลุ่มข้อมูลได้ดีกว่าวิธี SVM และจากการวิเคราะห์องค์ประกอบต่าง ๆ ของเว็บไซต์ทำให้รู้ว่าองค์ประกอบที่มีความสำคัญมากระหว่างเว็บไซต์ปกติและเว็บไซต์ไม่เหมาะสมนั้นคือองค์ประกอบของคำที่อยู่ภายใน BODY TAG อย่างไรก็ตามเทคนิคนี้ยังไม่ครอบคลุมกลุ่มคำที่มีความกำกวมและกรณีที่เว็บไซต์นั้นๆไม่สามารถตรวจสอบ HTML Code ได้

ภาควิชาคอมพิวเตอร์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร ปีการศึกษา 2552
ลายมือชื่อนักศึกษา.....
ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์ .....

48309322 : MAJOR : INFORMATION TECHNOLOGY
KEY WORD : WEB FILLTER / SUPPORT VECTOR MACHINE / PRINCIPAL COMPONENT ANALYSIS
CHANPAT PINUNRATCHATHORN : SYSTEM DEVELOPING OF WEB CONTENT FILTERING IN SECONDARY SCHOOL. THESIS ADVISOR : ASST.PROF. PANJAI TANTATSANAWONG, Ph.D. 154 pp.

This thesis has developed to filter inappropriate websites under the Linux operating system using Squid proxy software with ACL (Access Control List), which controlled of the web site accessing. The research applied two algorithms to analyze including: Principal Component Analysis (PCA) and Support Vector Machine (SVM). These algorithms are used for filtering websites to help control Internet access in schools and allowed students to receive the appropriate information.

Principal of the inappropriate websites filtering system used the TAG elements structure of the site as the following META, IMG, A HREF, SCRIPT, TITLE and BODY to create the filtering system. The inappropriate website brought to model the relationship and used PCA to determine the value of transformation for each component to create the models. Then model was tested by the site of 200 violence and drug websites, 200 pornographic web sites. The results of model evaluation showed that systems can filter inappropriate websites from normal websites, with accuracy for a group of violent and drug 89.5% and pornographic group 94.5%. Researcher also compared to the other filtering inappropriate algorithms called SVM using the same set of test data. The results showed that SVM method has the accuracy of filtering websites as the following: violent and drug 89%, pornographic 91%, while the PCA algorithms can segment data better than SVM. From the analysis showed that the elements in BODY TAG of various sites are very important to classify normal and inappropriate websites. However, these techniques do not apply to group words that are ambiguous and websites which can not check the HTML Code.

Department of Computing Graduate School, Silpakorn University Academic Year 2009
Student's signature .....
Thesis Advisor's signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ประสบความสำเร็จด้วยดีเป็นเพราะผู้วิจัยได้รับความกรุณาอย่างยิ่งจาก คณาจารย์ทุกท่านที่ประสิทธิ์ประสาทวิชาในการให้ความช่วยเหลือแนะนำและตรวจสอบแก้ไขข้อบกพร่องต่างๆ ของงานวิจัยด้วยความเอาใจใส่เป็นอย่างยิ่ง

ขอขอบพระคุณสมาชิกในครอบครัว “คุณพ่อและคุณแม่” ที่ให้การสนับสนุนช่วยเหลือให้กำลังใจด้วยความรักและความห่วงใย ทำให้ผู้วิจัยมีจิตใจที่มั่นคง มานะ บากบั่น เข้มแข็ง อดทน และมีกำลังใจในการเผชิญอุปสรรคต่างๆ จนทำให้งานวิจัยฉบับนี้สำเร็จลุล่วงได้ด้วยดี

ขอขอบพระคุณ ท่านเจ้าของเอกสารและงานวิจัยทุกท่าน ที่ผู้ศึกษาค้นคว้าได้นำมาอ้างอิงในการทำวิจัย ตลอดจนเพื่อนๆ พี่ ๆ น้องๆ ทุกท่าน ที่ให้ความช่วยเหลือ เอื้ออาทร พร้อมทั้งคำแนะนำที่เป็นประโยชน์ต่อการศึกษาค้นคว้า และคอยเป็นกำลังใจเสมอมาจนทำให้การศึกษาค้นคว้าครั้งนี้สำเร็จลุล่วงไปด้วยดี

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง
บทคัดย่อภาษาอังกฤษ ..... จ
กิตติกรรมประกาศ ..... ฉ
สารบัญตาราง ..... ฏ
สารบัญภาพ ..... ฑ
บทที่
1 บทนำ 1
ความเป็นมาและความสำคัญของปัญหา 1
วัตถุประสงค์การวิจัย 2
ขอบเขตของการวิจัย 2
ขั้นตอนในการสร้างและพัฒนาระบบ 3
ผลที่คาดว่าจะได้รับ 3
2 ทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง 4
สถิติ 4
ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) 4
วาเรียนซ์ (Variance) 5
โควาเรียนซ์ (Covariance) 5
โควาเรียนซ์เมตริกซ์ (Covariance Matrix) 6
ไอเกนแวลูส์และไอเกนเวกเตอร์ 7
หลักการวิเคราะห์แยกแยะส่วนประกอบ PCA (PCA : Principal Component Analysis) 9
ทฤษฎี SVM (SVM : Support Vector Machine) 11
กำหนดสมการสำหรับแบ่งข้อมูลออกเป็น 2 กลุ่ม (Binary Classification) 11
การกำหนดค่า Margin Maximization 13
ทฤษฎี Feature space 17

บทที่ หน้า

Kernel Functions 19
ขั้นตอนการวัดประสิทธิภาพของระบบ SVM 19
ขั้นตอนการติดต่อกับ Web Server 20
HTTP Return Code 21
การจำแนกประเภทของการกรองเว็บไซต์ที่ไม่เหมาะสม 22
ระบบการกรองเว็บไซต์ที่ไม่เหมาะสมที่เครื่อง Client 22
สร้างระบบการกรองเว็บไซต์ที่ไม่เหมาะสมที่ Gateway 22
วิธีป้องกันเว็บไซต์ที่ไม่เหมาะสมที่มีอยู่ในปัจจุบัน 22
Blacklists and Whitelists 23
Keyboard Blocking 23
Rating System 23
ระบบการป้องกันเว็บไซต์ที่ไม่เหมาะสมที่มีในปัจจุบัน 23
The Internet Content Rating Association (ICRA) 24
SafeSurf Rating 24
SmartFilter 24
Squidguard 24
Internet Access Content Management 24
DataReactor iMimic Networking Inc. 24
SITA URL filtering 25
WEB Filtering for WinProxy 25
SonicWALL Content Filtering Service (CFS) 25
FORTIGUARD WEB Filtering 25
ระบบพร็อกซี เซิร์ฟเวอร์(Proxy server) 25
การ Block เว็บไซต์โดยการสร้าง Blacklist 26
การ Block เว็บไซต์โดยใช้ Keyword 27
การทำ Transparency เพื่อให้ Client วิ่งเข้าใช้ Proxy อัตโนมัติ 27
ไฟล์บันทึกการใช้งาน access.log 28
อัลกอริทึมสำหรับการกรองเว็บไซต์ไม่เหมาะสม 28

บทที่ หน้า

Naive Bayes 28
K-Nearest Neighbor 29
Decision Tree 29
Support Vector Machines 29
Classification of hypertext data 29
Text Classification for hypertext filtering 29
Web Filtering Using Text Classification 29

3 วิธีดำเนินการวิจัย 30
กลุ่มตัวอย่างที่ใช้ในการวิจัย 30
เครื่องมือและอุปกรณ์ 30
โปรแกรมที่ใช้ในงานวิจัย 31
ขั้นตอนการศึกษาวิธีการเขียนโปรแกรมภาษาซีเพื่อติดต่อกับ Web Server 31
ศึกษาคุณลักษณะของเว็บไซต์ที่ไม่เหมาะสม 32
อัลกอริทึมสำหรับการสร้างระบบตรวจสอบเว็บไซต์ไม่เหมาะสม 32
สร้างระบบการกรองเว็บไซต์ไม่เหมาะสม 37
หลักการวิเคราะห์แยกแยะส่วนประกอบ PCA (PCA :Principal Component Analysis) 37
หลักการวิเคราะห์องค์ประกอบหลัก 38
การประเมินผล 39

4 ผลการดำเนินการวิจัย 40
เขียนโปรแกรมภาษาซีบน Ubuntu9.04 ในการติดต่อขอข้อมูล html code 40
วิเคราะห์หากลุ่มคำไม่เหมาะสมเพื่อใช้จำแนกเว็บไซต์ไม่เหมาะสมออกจากเว็บไซต์ปกติ 45
กลุ่มคำรุนแรง ยาเสพติด 45
กลุ่มคำลามกอนาจาร 51
วิเคราะห์หาจุดที่เหมาะสมที่ใช้ในการจำแนกเว็บไซต์ไม่เหมาะสมออกจากเว็บไซต์ปกติโดยใช้ทฤษฎี PCA (PCA : Principal Component Analysis) และ ทฤษฎีการคำนวณทางสถิติ 57

บทที่ หน้า

กำหนดมิติที่ใช้ในการทดสอบจากเนื้อหาในเว็บไซต์ 57
ตรวจสอบความแตกต่างของเนื้อหาเว็บไซต์ทั้ง 8 มิติ กับกลุ่มตัวอย่างทดลอง 58
วิเคราะห์จุดที่เหมาะสมในการแบ่งเว็บไซต์ลามกอนาจาร 59
วิเคราะห์จุดที่เหมาะสมในการแบ่งเว็บไซต์ความรุนแรง ยาเสพติด 69
ผลการทดสอบการแบ่งกลุ่มเว็บไซต์โดยใช้ทฤษฎี PCA (PCA : Principal Component Analysis) 72
การประเมินประสิทธิภาพจำนวนเว็บไซต์ที่ได้จากการคำนวณสมการเส้นตรง 73
ทดสอบวิเคราะห์องค์ประกอบ 3 มิติ 74
การประยุกต์ใช้ทฤษฎีอื่นในการแบ่งข้อมูล 75
สร้างระบบป้องกันการเข้าถึงเว็บไซต์ไม่เหมาะสม 78
การเตรียมความพร้อมสำหรับระบบ 78
การสร้างไฟล์ที่จำเป็นสำหรับการใช้งาน 79
วิเคราะห์การจำแนกเว็บไซต์โดยใช้ทฤษฎี SVM (SVM : Support Vector Machine) เพื่อเปรียบเทียบประสิทธิภาพการคัดกรอง 81
กำหนด Feature ที่ใช้ในการทดสอบ 81
ขั้นตอนก่อนการประมวลผล (pre-processing) 82
ขั้นตอนการประมวลผล (processing) 82
ผลการทดสอบการแบ่งกลุ่มเว็บไซต์โดยใช้ทฤษฎี SVM (SVM : Support Vector Machine) 83
ทำการเปรียบเทียบผลการทดสอบการคัดกรองเว็บไซต์ด้วยวิธี PCA (PCA : Principal Component Analysis) กับวิธี SVM (SVM : Support Vector Machine) 86
วิเคราะห์การเพิ่มกลุ่มคำไม่เหมาะสมแบบอัตโนมัติ 88
กรณีประเภทกลุ่มคำรุนแรง ยาเสพติด 88
กรณีประเภทกลุ่มคำลามกอนาจาร 89
การทดสอบประสิทธิภาพการทำงาน 91

บทที่ หน้า

ประสิทธิภาพด้านความถูกต้อง 91
ประสิทธิภาพด้านความเร็ว 92
ประสิทธิภาพในการรองรับเว็บ 2.0 97
5 สรุป อภิปรายผลและข้อเสนอแนะ 103
การบรรลุวัตถุประสงค์การวิจัย 103
ปัญหาและอุปสรรค 104
ข้อเสนอแนะ 105
ข้อสรุปใหม่ที่ได้จากการพัฒนาระบบการคัดกรองเว็บไซต์ 105
วิธีการใหม่ที่ได้จากการพัฒนาระบบการคัดกรองเว็บไซต์ 106
บรรณานุกรม 107

ภาคผนวก 110
ภาคผนวก ก คู่มือการใช้งานของระบบการตรวจสอบควบคุมเว็บไซต์ 111
ภาคผนวก ข คู่มือการใช้งาน SVM^{light} V6.02 121
ภาคผนวก ค รายชื่อเว็บไซต์ความรุนแรง ยาเสพติด 200 เว็บไซต์ 126
ภาคผนวก ง รายชื่อเว็บไซต์ลามกอนาจาร 150 เว็บไซต์ 134
ภาคผนวก จ รายชื่อเว็บไซต์ปกติ 200 เว็บไซต์ 141
ภาคผนวก ฉ กลุ่มคำลามกอนาจาร 112 คำ 149

ประวัติผู้วิจัย 154

สารบัญตาราง

ตารางที่		หน้า
1	ฟังก์ชั่นจัดการไฟล์ Config	32
2	ฟังก์ชั่นที่เกี่ยวข้องกับการจัดการ File Dicts	33
3	ฟังก์ชั่นที่เกี่ยวข้องกับการเก็บข้อมูลเว็บไซต์	33
4	ฟังก์ชั่นสำหรับทำกระบวนการ PCA	34
5	ฟังก์ชั่นที่ใช้จัดการโปรแกรม Squid	34
6	ฟังก์ชั่นที่เกี่ยวข้องกับการจัดการเว็บไซต์	35
7	ฟังก์ชั่นที่เกี่ยวข้องกับการทำงานของ Proxy-Server	35
8	อัตราส่วนจำนวนกลุ่มคำรุนแรง ยาเสพติดที่พบในเว็บไซต์ความรุนแรง ยาเสพติด ต่อ จำนวนกลุ่มคำรุนแรง ยาเสพติดที่พบในเว็บไซต์ปกติ	46
9	แสดงกลุ่มคำรุนแรง ยาเสพติดที่มีอำนาจการแจกแจงสูงสุด	50
10	อัตราส่วนจำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ลามกอนาจาร ต่อ จำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ปกติ	52
11	แสดงกลุ่มคำลามกอนาจารที่มีอำนาจการแจกแจงสูงสุด	56
12	แสดงคุณสมบัติของเนื้อหาเว็บไซต์	58
13	แสดงค่าเฉลี่ยคุณสมบัติของเว็บไซต์ลามกอนาจาร	59
14	แสดงค่า Eigenvalue ความแปรปรวน และความแปรปรวนรวม ที่คำนวณได้	64
15	แสดงประสิทธิภาพการคัดกรองเว็บไซต์ด้วยวิธี PCA (PCA : Principal Component Analysis)	72
16	แสดงประสิทธิภาพสมการเส้นตรงสำหรับการแบ่งประเภทเว็บไซต์ความรุนแรง ยาเสพติดกับเว็บไซต์ปกติ	73
17	แสดงประสิทธิภาพสมการเส้นตรงสำหรับการแบ่งประเภทเว็บไซต์ลามกอนาจารกับเว็บไซต์ปกติ	73
18	แสดงประสิทธิภาพการคัดกรองเว็บไซต์ด้วยวิธี SVM แบบ linear	84
19	แสดงประสิทธิภาพการคัดกรองเว็บไซต์ด้วยวิธี SVM แบบ Polynomial kernel	84
20	แสดงประสิทธิภาพการคัดกรองเว็บไซต์ด้วยวิธี SVM แบบ RBF kernel	85
21	แสดงประสิทธิภาพการคัดกรองเว็บไซต์ด้วยวิธี SVM แบบ Sigmoid kernel	86

ตารางที่ หน้า

22 แสดงผลการเปรียบเทียบประสิทธิภาพการคัดกรองเว็บไซต์ความรุนแรง ยาเสพติดด้วย วิธี PCA กับวิธี SVM 87

23 แสดงผลการเปรียบเทียบประสิทธิภาพการคัดกรองเว็บไซต์ลามกอนาจารด้วย วิธี PCA กับวิธี SVM 87

24 แสดงประสิทธิภาพด้านความถูกต้องของการจำแนกเว็บไซต์ 92

25 แสดงถึงเวลาที่ใช้ในการ Block เว็บไซต์(วินาที) ในช่วงของผู้ใช้งาน อินเทอร์เน็ตที่เพิ่มขึ้นเรื่อย 93

สารบัญภาพ

ภาพที่		หน้า
1	ขั้นตอนการวิเคราะห์องค์ประกอบหลัก PCA (PCA : Principal Component Analysis)	9
2	แสดงสมการเชิงเส้นที่ใช้แบ่งกลุ่มข้อมูล	12
3	แสดงเส้นแบ่งข้อมูลในระบบทั่ว ๆ ไป	12
4	แสดงเส้นแบ่งข้อมูลของซัพพอร์ตเวกเตอร์แมชชีน	13
5	แสดงค่า Margin ระหว่างจุดข้อมูลใด ๆ กับระนาบ Hyperplane	14
6	กรณีที่ข้อมูลสามารถแบ่งได้ด้วยสมการเชิงเส้น	17
7	กรณีที่ข้อมูลไม่สามารถแบ่งได้ด้วยสมการเชิงเส้น	18
8	การปรับ Feature Space เพื่อให้ข้อมูลสามารถแบ่งได้ด้วยสมการเชิงเส้น	18
9	การส่ง Request ไปยัง Web Server และการรับข้อมูลจาก Web Server	20
10	ขั้นตอนการติดต่อสื่อสารระหว่าง Client และ Server	21
11	การทำงานของ Proxy Server	26
12	ขั้นตอนการวิเคราะห์องค์ประกอบหลัก PCA (PCA :Principal Component Analysis)	37
13	ขั้นตอนการวิเคราะห์หากลุ่มคำไม่เหมาะสม ค่าทางสถิติ และทฤษฎี PCA (PCA : Principal Component Analysis)	45
14	กราฟแสดงอำนาจการแจงแจงของกลุ่มคำรุนแรง ยาเสพติด	50
15	กราฟแสดงอำนาจการแจงแจงของกลุ่มคำลามกอนาจาร	56
16	ตัวอย่างแสดงข้อมูลเว็บไซต์ลามกอนาจารที่มีการปรับค่าข้อมูล	60
17	แสดงค่าความแปรปรวนระหว่างมิติตัวเองกับมิติอื่น ๆ	61
18	แสดงผลกาคำนวณหาโควาเรียนซ์ทั้ง 8 มิติที่ปรับค่าแล้วของเว็บไซต์ลามกอนาจาร	61
19	แสดงค่า Eigenvalues ที่คำนวณได้	62
20	แสดงค่า Eigenvectors ที่คำนวณได้	63
21	แสดงค่า Eigenvectors ที่มีค่า Eigenvalues สูงสุด 2 อันดับ	65
22	แสดงตำแหน่งของข้อมูลแต่ละเว็บไซต์ของเว็บไซต์ลามกอนาจาร	66

ภาพที่ หน้า

23 ตัวอย่างแสดงข้อมูลเว็บไซต์ปกติที่มีการปรับค่าข้อมูล 67
24 แสดงตำแหน่งของข้อมูลแต่ละเว็บไซต์ของเว็บไซต์ปกติ 67
25 แสดงการแบ่งเว็บไซต์ลามกอนาจารออกจากเว็บไซต์ปกติ 68
26 ตัวอย่างแสดงข้อมูลเว็บไซต์ปกติที่ใช้ทดสอบความรุนแรงยาเสพติดที่มีการปรับ
ค่าข้อมูล..... 70
27 แสดงตำแหน่งของข้อมูลแต่ละเว็บไซต์ของเว็บไซต์ปกติ 70
28 แสดงการแบ่งเว็บไซต์ความรุนแรง ยาเสพติดออกจากเว็บไซต์ปกติ..... 71
29 แสดงระยะทางตามแนวแกน X Y และ Z เทียบกับจุด Origin 74
30 แสดงระนาบทั้ง 3 ระนาบบนระบบพิกัดฉาก 3 มิติ 74
31 แสดงการบอกพิกัดในระนาบ 3 มิติ..... 75
32 แสดงการแบ่งกลุ่มข้อมูลโดยทฤษฎี K-means..... 77
33 แสดงการจัดรูป Model มิติที่ลดรูปแล้วด้วยเวกเตอร์ 77
34 แสดงผลการคัดกรองเว็บไซต์กรณีกลุ่มคำเพิ่มขึ้น 10 คำ..... 90
35 แสดงผลการคัดกรองเว็บไซต์กรณีกลุ่มคำเพิ่มขึ้น 100 คำ..... 91
36 กราฟแสดงประสิทธิภาพความเร็วของระบบ 95
37 กราฟแสดงค่าเฉลี่ยของเวลาในการเข้าถึงเว็บไซต์..... 96
38 กราฟแสดงส่วนเบี่ยงเบนมาตรฐานของเวลาในการเข้าถึงเว็บไซต์..... 96
39 กลุ่มตัวอย่าง Web 2.0 เว็บไซต์ปกติ สำหรับการทดสอบเว็บไซต์ลามกอนาจาร .. 98
40 แสดงถึงระบบสามารถตรวจสอบ WEB 2.0 ประเภทลามกอนาจารได้ 99
41 กลุ่มตัวอย่าง Web 2.0 เว็บไซต์ปกติสำหรับการทดสอบเว็บไซต์ความรุนแรง 100
ยาเสพติด .....
42 แสดงถึงระบบสามารถตรวจสอบ WEB 2.0 ประเภทความรุนแรง ยาเสพติดได้..... 101
43 กลุ่มตัวอย่าง Web 2.0 เว็บไซต์ลามกอนาจารสำหรับการทดสอบเว็บไซต์ลามก 101
อนาจาร.....
44 แสดงถึงระบบสามารถตรวจสอบ WEB 2.0 ประเภทลามกอนาจารได้ 102
45 หน้าจอ Login 115
46 หน้าจอแจ้งข่าวสารในหน้า Home ผ่าน Menu News..... 115
47 หน้าจอเมนู Config Files..... 117

ภาพที่ หน้า

48 หน้าจอเมนู Config ProxyFilter 117
49 หน้าจอเพิ่มกลุ่มคำไม่เหมาะสม 118
50 หน้าจอแสดงกลุ่มคำไม่เหมาะสม 119
51 หน้าจอแสดงกลุ่มคำไม่เหมาะสมเพิ่มเติม 120
52 แสดงรูปแบบข้อมูล Data Train 122
53 ตัวอย่างการสร้าง model ด้วยวิธี SVM แบบ linear 123
54 ตัวอย่างการทดสอบข้อมูลด้วยวิธี SVM แบบ linear 124
55 ตัวอย่าง output ที่ได้จากการทดสอบ 125

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

เนื่องจากปัจจุบัน Internet มีส่วนสำคัญมากกับการตัดสินใจของวัยรุ่น โดยวัยรุ่นส่วนใหญ่จะหาความรู้จากทาง Internet เพื่อประกอบกับการตัดสินใจในเรื่องต่างๆ ซึ่งใน Website แต่ละที่ก็จะมีเนื้อหาที่แตกต่างกันออกไป เช่น เว็บไซต์ที่ให้ความรู้ความบันเทิง และข่าวสารข้อมูล แต่ก็มี Website เป็นจำนวนมากที่ทำขึ้นมาแล้ว ไม่ได้มีเนื้อหาในทางที่เป็นประโยชน์ต่อสังคม เด็กและเยาวชน ซึ่งเว็บเหล่านี้ทำให้เกิดปัญหาทางสังคมต่างๆ ตามมาได้ในภายหลัง เช่น เว็บไซต์ลามก อนาจาร ขายของที่ผิดกฎหมายผ่าน Internet เว็บไซต์เกี่ยวกับยาเสพย์ติด และ การฆ่าตัวตาย ทำให้วัยรุ่นอยากรู้อยากลอง นำไปปฏิบัติตาม หมกมุ่น หรืออาจส่งผลเสียต่อสภาวะจิตใจ โดยที่พ่อแม่ของบุตรหลานไม่สามารถที่จะควบคุมได้

ปัจจุบันโรงเรียนหรือสถานที่ศึกษาต่าง ๆ ได้มีการผลักดันให้นักเรียน ค้นคว้าหาความรู้ทาง Internet มากขึ้น เพราะ สามารถหาความรู้ได้รวดเร็ว จาก Website ต่าง ๆ แต่ก็มี Website อีกมากมายที่ไม่ได้ทำขึ้นมาเพื่อให้ความรู้ทางการศึกษาของเด็กเท่านั้น แต่กลับให้โทษมากกว่าตามที่ได้กล่าวมาแล้ว ซึ่งเป็นการยากที่จะห้ามนักเรียนไม่ให้เข้าไปดู หรือ ใช้งาน Internet ไปในทางที่ไม่เหมาะสม ถึงแม้ว่าปัจจุบันทางภาครัฐได้มีโครงการ Cyber Inspector เพื่อทำการตรวจสอบเว็บไซต์ต่างๆ ที่มีอยู่ในปัจจุบันว่าเหมาะสมหรือไม่ โดยเว็บไซต์ไหนที่ไม่เหมาะสมก็จะถูก block ซึ่งเป็นวิธีการที่ดีอย่างหนึ่ง แต่ก็ไม่อาจทั่วถึง เนื่องจากในทุกวันนี้มี Website ที่เกิดขึ้นมาใหม่เป็นจำนวนมากทุกวัน

Internet ในโรงเรียนต่าง ๆ นั้น โดยส่วนใหญ่แล้วจะมีการตั้งเครื่อง Server ขึ้นมา และใช้ระบบปฏิบัติการต่าง ๆ มาทำเป็น Server เพื่อช่วยเพิ่มความสามารถด้านต่าง ๆ ให้ดีขึ้น เช่น มีการ Block website ที่ไม่เหมาะสม, มีการทำ Cache เพื่อช่วยให้ใช้ Internet ได้เร็วขึ้น อีกทั้งทางภาครัฐเองยังมีนโยบายสนับสนุนให้นำ ระบบปฏิบัติการ Linux ซึ่งเป็น Open source มาทำเป็นระบบปฏิบัติการในโรงเรียนอีกด้วย

ตัวระบบปฏิบัติการ Linux จะมีโปรแกรมที่ชื่อว่า Squid ติดมาด้วย ซึ่งในตัวโปรแกรมจะมีการตั้งกฎการใช้งานต่าง ๆ ของ Internet เรียกว่าACL (Access Control List) เพื่อให้ทาง

โรงเรียนควบคุมความเหมาะสมในการใช้งาน Internet ของนักเรียน และสามารถตั้งกฎหรือข้อบังคับในการใช้งานได้ โดยจะนำมาใช้ในการควบคุมการใช้ Internet เช่น ห้าม Download file, ห้ามเข้าดู Website ที่ไม่เหมาะสม หรือให้ใช้งาน Internet เฉพาะบางเวลา เป็นต้น

งานวิจัยนี้จึงนำเสนอการควบคุมการใช้งาน Website ที่ไม่เหมาะสมสำหรับนักเรียนในระดับมัธยมศึกษา เพื่อให้นักเรียนในโรงเรียนต่างๆได้ท่องเว็บไซต์อย่างปลอดภัยในระดับหนึ่ง โดยการสร้างระบบคัดกรองWebsite ขึ้นมา และใช้โปรแกรม Squid ซึ่งจะมี ACL (Access Control List) เป็นตัวควบคุมการใช้งาน Website ของเด็กนักเรียน และได้มีการพัฒนาระบบคัดกรอง Website เพื่อให้การควบคุม ACL (Access Control List) เป็นไปโดยอัตโนมัติ

วัตถุประสงค์การวิจัย

1. เพื่อศึกษาวิธีการคัดกรองเนื้อหาของเว็บไซต์ที่ไม่เหมาะสม
2. เพื่อพัฒนาระบบคัดกรองเนื้อหาของเว็บไซต์
3. เพื่อประเมินผลระบบที่พัฒนาขึ้น

ขอบเขตของการวิจัย

1. ศึกษาการทำงานของ โปรแกรม Squid-2.7 stable3 โดยใช้ ACL (ACL : Access Control List) ในการควบคุมการใช้งาน Internet บนเครื่องแม่ข่าย

2. ศึกษาการใช้งาน Website ภายในโรงเรียน ระดับมัธยมศึกษาตอนต้น และมัธยมศึกษาตอนปลาย

3. พัฒนาส่วนของการวิเคราะห์เนื้อหาบนเว็บไซต์ ใช้เฉพาะ HTML Code ในการจำแนกเว็บไซต์

4. ศึกษาเว็บไซต์ที่มีเนื้อหาเกี่ยวกับ ความรุนแรง ยาเสพติด และ ลามกอนาจาร

5. ศึกษาการคัดแยกข้อมูล โดยใช้ Text Processing

6. สนับสนุนการทำงานร่วมกับโปรโตคอล HTTP เท่านั้น

7. ระบบสามารถทำงานร่วมกับ Web Browser ที่เป็น Internet Explorer และ Firefox

8. ทำการทดสอบกับกลุ่มตัวอย่าง ซึ่งใช้เว็บไซต์ ที่ทำการทดสอบเกี่ยวกับ ความรุนแรง ยาเสพติด จำนวน 200 เว็บไซต์ เกี่ยวกับ ลามกอนาจารจำนวน 150 เว็บไซต์

9. ทำการตรวจสอบความเหมาะสมโดยคิดเป็นเปอร์เซ็นต์ของประสิทธิภาพ และความผิดพลาด

10. การบล็อกเว็บไซต์เป็นการบล็อกในแต่ละหน้าของเว็บไซต์

11. อ้างอิง Squid-2.7 Stable3 ที่ทำหน้าที่เป็น Proxy Cache Server ทำงานบน Linux Ubuntu9.04

12. ทำการทดสอบกับ เว็บไซต์ที่ใช้รูปแบบภาษา UTF-8 เท่านั้น

ขั้นตอนในการสร้างและพัฒนาระบบ

1. ศึกษาการทำงานของระบบ Server ที่ใช้ระบบปฏิบัติการ Linux Ubuntu9.04

2. ศึกษาการทำงานของโปรแกรม squid โดยใช้ ACL (Access Control List) ในการควบคุมการใช้งาน Internet

3. เก็บรวบรวมข้อมูลจากเอกสาร และแหล่งข้อมูลที่เกี่ยวข้อง และลักษณะของการบันทึกการใช้งานของเครื่องแม่ข่ายที่ทำหน้าที่เป็น Proxy Cache Server

4. ศึกษาอัลกอริทึม PCA (PCA : Principal Component Analysis) ที่จะนำมาพิจารณาในส่วนของ Text Detection

5. ศึกษาการทำงานของ Access.log เพื่อนำมาเป็นข้อมูลในการพิจารณาการใช้งานเว็บไซต์ของนักเรียนภายในโรงเรียน

6. เขียนโปรแกรมและทำการทดลองกับระบบงานจริง

7. ทำการทดสอบกับกลุ่มตัวอย่าง ซึ่งใช้ เว็บไซต์ ที่ทำการทดสอบ เกี่ยวกับ ลามกอนาจาร จำนวน 150 เว็บไซต์ เกี่ยวกับความรุนแรง ยาเสพติด จำนวน 200 เว็บไซต์

8. ทำการทดสอบความเหมาะสมโดย คิดเป็น เปอร์เซ็นต์ประสิทธิภาพ และความผิดพลาด

9. สรุปผลการดำเนินงาน

10. จัดทำรายงานวิทยานิพนธ์

ผลที่คาดว่าจะได้รับ

1. ระบบที่ใช้ควบคุม ความเหมาะสมของเนื้อหาบนเว็บไซต์

2. ช่วยลดภาระของบุคลากรที่เกี่ยวข้องในการควบคุมการใช้งาน Website ของนักเรียน

3. ช่วยควบคุมการใช้งาน Internet ในโรงเรียนให้นักเรียนได้รับข้อมูลที่เหมาะสม

4. ช่วยลดพฤติกรรมการเลียนแบบ Website ที่ไม่เหมาะสมส่งผลให้เกิดความเดือดร้อนต่อครอบครัว และสังคม

บทที่ 2

ทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง

การศึกษาทฤษฎีและผลงานวิจัยที่เกี่ยวข้องในบทนี้ผู้วิจัยได้ศึกษาจากเอกสารแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง ประกอบด้วย

1. สถิติ
2. PCA (PCA : Principal Components Analysis)
3. SVM (SVM : Support Vector Machine)
4. ขั้นตอนการติดต่อกับ Web Server
5. การจำแนกประเภทของการกรองเว็บไซต์ที่ไม่เหมาะสม
6. วิธีป้องกันเว็บไซต์ที่ไม่เหมาะสมที่มีอยู่ในปัจจุบัน
7. ระบบป้องกันเว็บไซต์ที่ไม่เหมาะสมที่มีในปัจจุบัน
8. ระบบพร๊อกซี่ เซิร์ฟเวอร์ (Proxy server)
9. อัลกอริทึมสำหรับการกรองเว็บไซต์ไม่เหมาะสม

1. สถิติ (Lindsay, 2002)

สถิติสามารถนำมาใช้เมื่อมีชุดข้อมูลขนาดใหญ่และต้องการวิเคราะห์ความสัมพันธ์ระหว่างแต่ละจุดของชุดข้อมูลนั้น ในที่นี้จะกล่าวถึงการวัดที่กระทำกับกลุ่มของข้อมูล เช่น ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation), วาเรียนซ์ (Variance), โควาเรียนซ์ (Covariance) และ โควาเรียนซ์เมตริกซ์ (Covariance Matrix) ซึ่งเป็นสถิติที่จะนำมาใช้ใน PCA (PCA : Principal Components Analysis)

1.1 ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation)

ค่าเบี่ยงเบนมาตรฐานของชุดข้อมูลเป็นการวัดการกระจายตัวของข้อมูล คำนิยามของค่าเบี่ยงเบนมาตรฐานคือระยะทางเฉลี่ยจากค่ากลางของชุดข้อมูลถึงจุดข้อมูลหนึ่งๆ การคำนวณค่าเบี่ยงเบนมาตรฐานมีสมการดังนี้

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}} \tag{2.1}$$

การคำนวณหาค่าเบี่ยงเบนมาตรฐานสามารถหาได้จากกำลังสองของระยะทางจากแต่ละจุดของข้อมูลไปยังค่าเฉลี่ยของกลุ่มข้อมูล ทำการรวมค่าที่คำนวณได้ทั้งหมดหารด้วยจำนวนสมาชิกของกลุ่มข้อมูล (n) ลบด้วย 1 แล้วทำการหาค่ารากที่สอง

1.2 วาเรียนซ์ (Variance)

วาเรียนซ์เป็นตัววัดการกระจายตัวของข้อมูลอีกแบบหนึ่งมีลักษณะแบบเดียวกับค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ซึ่งมีสมการดังนี้

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)} \tag{2.2}$$

ซึ่งอาจกล่าวได้ว่าวาเรียนซ์เป็นกำลังสองของค่าเบี่ยงเบนมาตรฐาน (s^2) โดยทั้งค่าเบี่ยงเบนมาตรฐานและวาเรียนซ์ต่างก็เป็นการวัดการกระจายตัวของข้อมูลแต่ ค่าเบี่ยงเบนมาตรฐานจะนิยมใช้มากกว่า

1.3 โควาเรียนซ์ (Covariance)

จากที่กล่าวมาค่าเบี่ยงเบนมาตรฐานและวาเรียนซ์เป็นการวัดการกระจายตัวของข้อมูล 1 มิติ อย่างไรก็ตามมีชุดข้อมูลจำนวนมากที่มีมากกว่า 1 มิติ และจุดประสงค์ของการวิเคราะห์ข้อมูลทางสถิติเพื่อหาว่าข้อมูลมีความสัมพันธ์ระหว่างมิติหรือไม่ โดยปกติแล้วโควาเรียนซ์จะใช้กับข้อมูล 2 มิติ ถ้าทำการคำนวณโควาเรียนซ์ระหว่างข้อมูล 1 มิติและตัวมันเองจะได้ค่าวาเรียนซ์ แต่ถ้าข้อมูลมี 3 มิติ (x,y,z) สามารถคำนวณค่าโควาเรียนซ์ได้โดยคำนวณโควาเรียนซ์ระหว่าง x กับ y , x กับ z และ y กับ z การคำนวณค่าโควาเรียนซ์ระหว่าง x กับ x , y กับ y และ z กับ z จะได้ค่าวาเรียนซ์ของ x, y และ z

สมการของโควาเรียนซ์คล้ายกับสมการของวาเรียนซ์ ซึ่งสมการของวาเรียนซ์เป็นดังนี้

$$\mathrm{var}(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)} \tag{2.3}$$

และสมการของโควาเรียนซ์เป็นดังนี้

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)} \tag{2.4}$$

ถ้าค่าโควาเรียนซ์ที่คำนวณออกมามีค่าเป็นบวกแสดงว่าค่าข้อมูลมีการเพิ่มขึ้นทั้งสองมิติ แต่ถ้าค่าโควาเรียนซ์เป็นลบแสดงว่าข้อมูลมีค่าเพิ่มขึ้นหนึ่งมิติ ส่วนอีกมิติมีค่าลดลง แต่ถ้าค่าเป็นศูนย์แสดงว่าข้อมูลทั้งสองมิติไม่ได้ขึ้นต่อกัน โดยค่าของ cov(X,Y) จะมีค่าเท่ากับ cov(Y,X)

1.4 โควาเรียนซ์เมตริกซ์ (Covariance Matrix)

จากที่ได้กล่าวมาข้างต้นโดยโควาเรียนซ์จะคำนวณข้อมูล 2 มิติ ถ้ามีข้อมูลมากกว่า 2 มิติ แสดงว่ามีการคำนวณโควาเรียนซ์มากกว่าหนึ่งตัว เช่น ชุดข้อมูลที่มี 3 มิติ (x, y, z) สามารถคำนวณ cov(x,y), cov(x,z) และ cov(y,z) ถ้ามีข้อมูล n มิติ จะสามารถคำนวณค่าโควาเรียนซ์ได้แตกต่างกัน

$$\frac{n!}{(n-2)!*2}$$

ทางที่ง่ายสำหรับการคำนวณค่าโควาเรียนซ์ระหว่างมิติหลายๆมิติ ทำได้โดยการนำค่า โควาเรียนซ์ทั้งหมดใส่ลงในเมตริกซ์ ซึ่งนิยามสำหรับโควาเรียนซ์เมตริกซ์สำหรับชุดข้อมูลที่มี n มิติ เป็นดังนี้

$$C^{nxn} = (c_{i,j}, c_{i,j} = \text{cov}(Dim_i, Dim_j)) \tag{2.5}$$

ซึ่ง C^{nxn} คือเมตริกซ์ที่มี n แถว n คอลัมภ์ และข้อมูลในเมตริกซ์คือผลลัพธ์ของการคำนวณโควาเรียนซ์ระหว่างมิติ 2 มิติที่ต่างกัน เช่น ข้อมูลที่อยู่ในแถว 2 คอลัมภ์ 3 คือการคำนวณโควาเรียนซ์ระหว่างมิติที่ 2 และมิติที่ 3

ตัวอย่างเช่น ข้อมูลมี 3 มิติ (x,y,z) ดังนั้นโควาเรียนซ์เมตริกซ์จะมี 3 แถวและ 3 คอลัมภ์ ดังนี้

$$C = \begin{pmatrix} \text{cov}(x,x) & \text{cov}(x,y) & \text{cov}(x,z) \\ \text{cov}(y,x) & \text{cov}(y,y) & \text{cov}(y,z) \\ \text{cov}(z,x) & \text{cov}(z,y) & \text{cov}(z,z) \end{pmatrix} \tag{2.6}$$

พิจารณาจากโควาเรียนซ์เมตริกซ์ จะเห็นว่าตามแนวเส้นทะแยงมุมบนซ้ายไปยังล่างขวา เป็นค่าโควาเรียนซ์ระหว่างมิตินั้นและตัวมันเอง หรือก็คือวาเรียนซ์ของมิตินั้น และจาก cov(x,y) มีค่าเท่ากับ cov(y,x) จะเห็นได้ว่าเมตริกซ์นี้ สมมาตรบนเส้นทแยงมุมนี้ด้วย

1.5 ไอเกนแวลูส์และไอเกนเวกเตอร์ (Eigenvalue eigenvector and eigenspace, 2010)

จากทฤษฎีความสัมพันธ์ไอแกนแวลูส์ซึ่งสัมพันธ์กับไอแกนเวกเตอร์ สามารถพบไอแกนแวลูส์และ ไอแกนเวกเตอร์มาเป็นคู่นั้น หมายถึง เมื่อหาไอแกนแวลูส์ได้ก็จะได้ไอแกนเวกเตอร์ด้วย โดยความสัมพันธ์ของไอแกนแวลูส์ และไอแกนเวกเตอร์ เป็นดังสมการต่อไปนี้

$$AX = \lambda IX \tag{2.7}$$

โดย A คือ เมตริกซ์ที่จำนวนคอลัมภ์และจำนวนแถวเท่ากัน

λ คือ เลขจำนวนจริงหรือจำนวนเชิงซ้อน

X คือ เวกเตอร์

I คือ เมตริกซ์เอกลักษณ์ของ A (Identify matrix)

จากสมการ 1 จะได้ว่า λ จะถูกเรียกว่า ไอแกนแวลูส์ของ A เวกเตอร์ X จะถูกเรียกว่า ไอแกนเวกเตอร์ของ A และสามารถเขียนสมการ ใหม่ได้ดังนี้

$$AX - \lambda IX = 0 \tag{2.8}$$

$$(A - \lambda I)X = 0 \tag{2.9}$$

เราสามารถสรุปจากคุณสมบัติดีเทอมิเนนท์(characteristic determinant)ได้ดังนี้

$$\det(A - \lambda I) = 0 \tag{2.10}$$

สามารถคำนวณหา λ ซึ่งเป็น ไอแกนแวลูส์ ของ A ได้ n ค่า (n คือจำนวนมิติของ A) จากสมการ (2.10) เมื่อได้ λ แทนค่าในสมการ (2.8) จะสามารถคำนวณหา X ซึ่งเป็น ไอแกนเวกเตอร์ของ A ได้ n เวกเตอร์เช่นกัน

จากทฤษฎี

ถ้า เมตริกซ์ A = $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

จากสมการที่ 2.10 หาค่า ไอแกนแวลูส์ได้ดังนี้

$$\det\left(\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix} - \begin{bmatrix}\lambda & 0\\ 0 & \lambda\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \right) = 0$$

$$\det\left(\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix} - \begin{bmatrix}\lambda & 0\\ 0 & \lambda\end{bmatrix} \right) = 0$$

$$\det\begin{bmatrix}2-\lambda & 1\\ 1 & 2-\lambda\end{bmatrix} = (2-\lambda)^2 - 1 = 0$$

รากของสมการนี้คือ $\lambda = 1$ และ $\lambda = 3$

ดังนั้น ไอเกนแวลูส์ $= \begin{bmatrix}1 & 0\\ 0 & 3\end{bmatrix}$

การหาไอเกนเวกเตอร์จากสมการที่ 1

กรณี $\lambda = 3$ $\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}\begin{bmatrix}X\\ Y\end{bmatrix} = 3\begin{bmatrix}X\\ Y\end{bmatrix}$

$$2X + Y = 3X$$

$$X + 2Y = 3Y$$

ดังนั้น $X = Y$

ถ้า $Y = 1$ จะได้ไอเกนเวกเตอร์ $= \begin{bmatrix}1\\ 1\end{bmatrix}$

กรณี $\lambda = 1$ $\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}\begin{bmatrix}X\\ Y\end{bmatrix} = 1\begin{bmatrix}X\\ Y\end{bmatrix}$

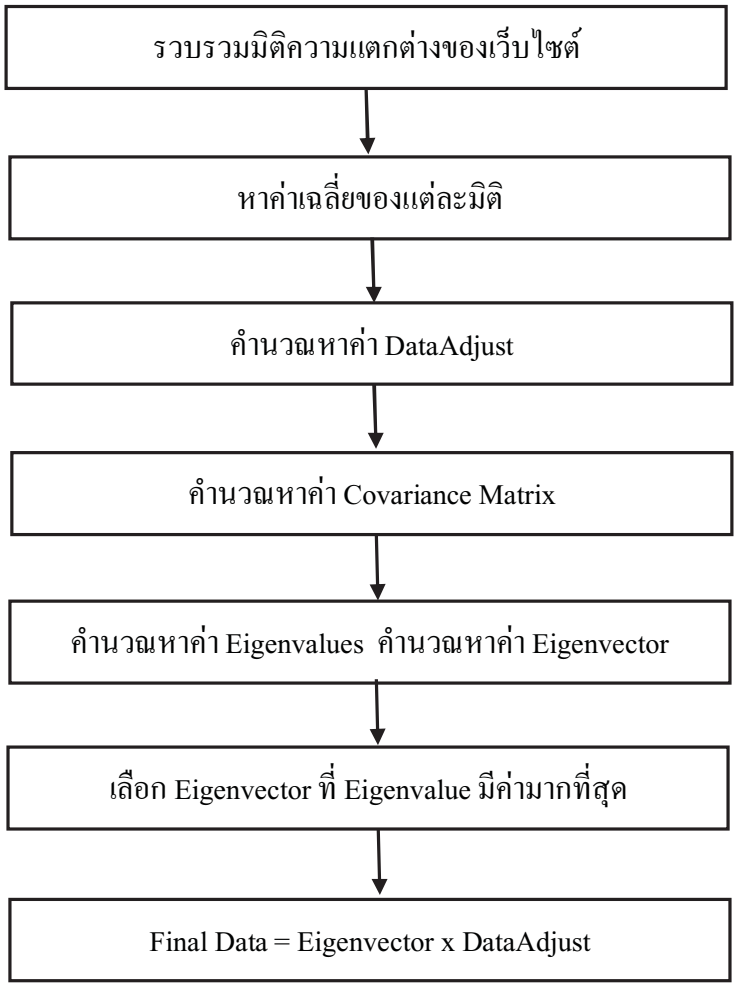
$$2X + Y = X$$

$$X + 2Y = Y$$

ดังนั้น $X = -Y$

ถ้า $Y = -1$ จะได้ไอเกนเวกเตอร์ $\begin{bmatrix}1\\ -1\end{bmatrix}$

2. หลักการวิเคราะห์แยกแยะส่วนประกอบ PCA (PCA : Principal Component Analysis)



ภาพที่ 1 ขั้นตอนการวิเคราะห์องค์ประกอบหลัก PCA (PCA : Principal Component Analysis)

หลักการวิเคราะห์แยกแยะส่วนประกอบ PCA (PCA : Principal Component Analysis) (พิชานี ทัศนเสถียร และภักดิ์ทูล ใจทอง 2549 : 26-28) คือ วิธีในการบ่งชี้ถึงรูปแบบของข้อมูลและช่วยแสดงข้อมูลให้เห็นถึงจุดเด่นของความเหมือนและความต่างของข้อมูล อย่างไรก็ตามรูปแบบของข้อมูลนั้นยากที่จะหาได้ เมื่อข้อมูลนั้นมีหลายมิติ (Dimension) และยังยากที่จะแสดงเป็นรูปภาพหรือกราฟได้ ดังนั้น PCA จึงเป็นเครื่องมือที่มีประสิทธิภาพอย่างมากในการแยกแยะหรือจำแนกข้อมูล

นอกจากนี้ PCA ยังมีประโยชน์ในการช่วยกำหนดรูปแบบของข้อมูล และบีบอัดข้อมูลได้โดยการ ลดจำนวนมิติ ปราศจากการสูญเสียเนื้อหาของข้อมูล โดยการวิเคราะห์แยกแยะส่วนประกอบการใช้งานมีวิธีการทำดังนี้

ขั้นตอนที่ 1 รวบรวมข้อมูล รวบรวมข้อมูลโดยพิจารณาถึงมิติ ของปัจจัย

ขั้นตอนที่ 2 หาค่า (Mean) หาค่ากลางของข้อมูลแต่ละมิติที่ได้รวบรวมไว้

ขั้นตอนที่ 3 หาค่า DataAdjust โดยนำข้อมูลแต่ละมิติหักลบออกจากค่ากลาง (Mean) ของแต่ละมิติ เราจะต้องหักลบค่ากลางออกจากแต่ละข้อมูลของมิติ ค่ากลางที่จะทำการลบนี้เป็นค่าเฉลี่ยตามขวางของแต่ละมิติ ดังนั้น ค่า x จะต้องลบด้วย $\bar{x}$ และค่า y จะต้องลบด้วย $\bar{y}$ ทำให้ผลลัพธ์ของชุดข้อมูลที่มีค่ากลางเท่ากับ 0 (ข้อมูลที่ถูกลบออกจากค่ากลางแล้ว จะเรียกว่า Data Adjust)

ขั้นตอนที่ 4 การคำนวณค่าเมทริกซ์โควาเรียนซ์ (Covariance Matrix) สมมติว่าหากข้อมูลเป็น 2 มิติ ทำให้เมทริกซ์โควาเรียนซ์เป็น 2 x 2

ขั้นตอนที่ 5 การคำนวณไอแกนเวกเตอร์ (Eigenvectors) และไอแกนแวลูส์ (Eigenvalues) ของค่า เมทริกซ์โควาเรียนซ์ เมื่อค่าเมทริกซ์โควาเรียนซ์เป็น สี่เหลี่ยมจัตุรัสแล้ว(จำนวน Row เท่ากับจำนวนของ Column) เราจะสามารถคำนวณค่าของไอแกนเวกเตอร์ และไอแกนแวลูส์ สำหรับเมทริกซ์ได้ ดังนั้นในกระบวนการทำงานของการสร้างไอแกนเวกเตอร์ของเมทริกซ์โควาเรียนซ์ เราจึงสามารถแยกเส้นที่อยู่บนกราฟซึ่งเป็นลักษณะเด่นของข้อมูลออกมาได้

ขั้นตอนที่ 6 เลือก Eigenvector ที่ Eigenvalues มีค่ามากที่สุด จะสังเกตได้ว่าไอแกนแวลูส์จะเป็นค่าที่ค่อนข้างแตกต่างกัน ในความเป็นจริงไอแกนเวกเตอร์กับค่า ไอแกนแวลูส์ที่มีค่าสูงสุด จะเป็นส่วนประกอบที่สำคัญของชุดข้อมูล เช่น หากไอแกนเวกเตอร์ กับไอแกนแวลูส์ที่มีค่าสูง จะเป็นจุดที่อยู่ตรงกลางของชุดข้อมูล ซึ่งเป็นความสัมพันธ์ที่เห็นได้อย่างชัดเจนระหว่างมิติของข้อมูลกับไอแกนเวกเตอร์ที่ถูกพบจากเมทริกซ์โควาเรียนซ์ ในขั้นตอนต่อไปไอแกนแวลูส์จะถูกเรียงเป็นลำดับโดย จากค่าสูงสุดไปต่ำสุด ทำให้สามารถเรียงลำดับความสำคัญของปัจจัยได้ ดังนั้นจึงสามารถตัดสินใจที่จะเพิกเฉยต่อ ปัจจัยที่มีความสำคัญน้อย หรือไอแกนแวลูส์มีค่าน้อย ซึ่งไม่ถือว่าเป็นการสูญเสียข้อมูลมากนัก โดยสามารถตัดบางปัจจัยออกไปได้ ซึ่งข้อมูลสุดท้ายจะมีมิติน้อยกว่าข้อมูลเริ่มต้น ถ้าข้อมูลเริ่มต้นมี n มิติ และ ทำการคำนวณ n ของไอแกนเวกเตอร์ และไอแกนแวลูส์ ดังนั้นเราจะเลือกเพียงไอแกนเวกเตอร์p แรก และข้อมูลชุดสุดท้าย จะมีเพียง p มิติ จากนั้นต้องทำการแปลงให้อยู่ในรูปของเวกเตอร์ (Feature vector)โดยการนำค่าไอแกนเวกเตอร์ที่เลือกมาจาก list ของไอแกนเวกเตอร์ และเปลี่ยนเป็นเมตริกซ์ให้ไอแกนเวกเตอร์อยู่ในคอลัมน์

$$\text{Eigenvector} = (\text{eig1 eig2 eig3} \ldots \text{eign}) \qquad (2.11)$$

ขั้นตอนที่ 7 การคำนวณหาค่า FinalData จะเป็นขั้นตอนสุดท้ายของ PCA และเป็นขั้นตอนที่ง่ายที่สุด โดยเราจะเลือกส่วนประกอบ หรือปัจจัยที่ดีที่สุด (eigenvector) นำมา Dot Matrix กับ DataAdjust เพื่อให้ได้ FinalData ที่เป็นข้อมูลตามมิติที่เรากำหนด

$$\text{FinalData} = \text{Eigenvector} \times \text{DataAdjust} \tag{2.12}$$

3. ทฤษฎี SVM (SVM : Support Vector Machine) (ธีรพงศ์ โหมดหิรัญ 2548 : 37-42)

วิธีการของซัพพอร์ตเวกเตอร์แมชชีน หรือ SVM จัดเป็นเทคนิคที่ใช้ในการแก้ไขปัญหาด้านการแบ่งกลุ่มข้อมูล (Classification) โดยอาศัยหลักการทางคณิตศาสตร์เพื่อหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มจากข้อมูลที่ถูกป้อน (Pattern) เพื่อใช้ในกระบวนการแบ่งกลุ่มโดยจุดประสงค์ของ SVM คือการสร้างระนาบที่สามารถแบ่งชุดข้อมูลได้ดีที่สุด (Optimal Separating Hyperplane)

สมมติให้มีชุดข้อมูลอยู่หนึ่งกลุ่มในการสร้างเส้นแบ่งข้อมูลสามารถแบ่งข้อมูลออกเป็น 2 กลุ่ม และมีคำตอบที่เป็นไปได้คือ $y = \{-1, 1\}$ โดยค่า y นี้จะเป็นผลลัพธ์ที่ต้องการให้ SVM เรียนรู้เพื่อใช้ในการแบ่งกลุ่มของข้อมูล

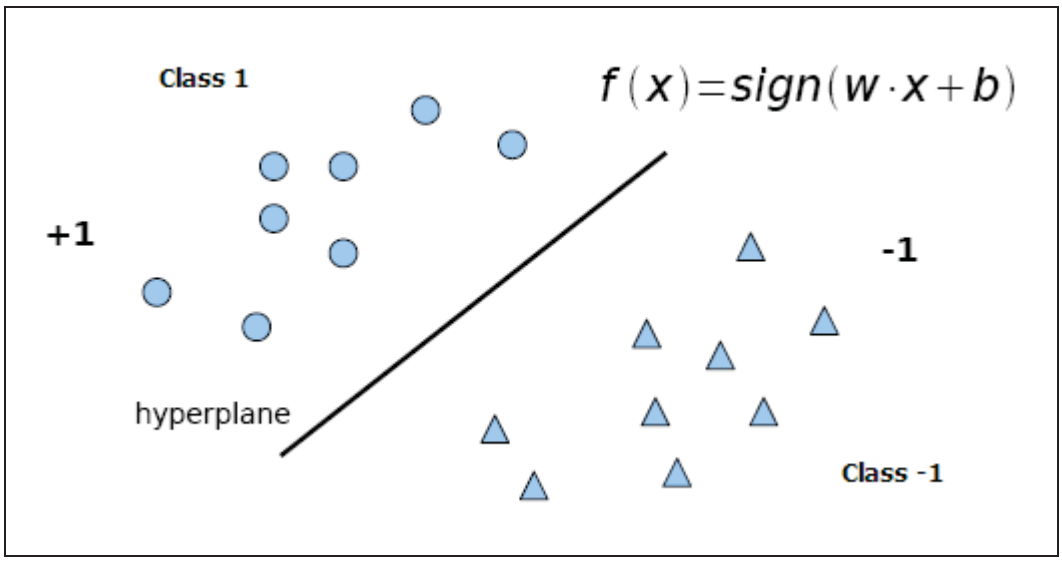
สำหรับพื้นฐานเดิมของ SVM นั้นถูกนำมาใช้กับข้อมูลที่ระนาบระหว่าง 2 กลุ่มเป็นลักษณะเชิงเส้น (Linear) แต่ในความเป็นจริงบางครั้งข้อมูลที่นำมาอาจจะมีระนาบแบ่งกลุ่มที่ไม่เป็นเชิงเส้น ซึ่งสามารถปรับ SVM ให้ใช้กับปัญหานี้ได้โดยการใช้เคอร์เนลฟังก์ชัน (Kernel Function) มาเปลี่ยน Feature Space ทฤษฎีซัพพอร์ตเวกเตอร์แมชชีน หรือ SVM (SVM : Support Vector Machine) มีหลักการที่สำคัญดังนี้

3.1 กำหนดสมการสำหรับแบ่งข้อมูลออกเป็น2 กลุ่ม (Binary Classification)

ถ้าเราเทียบกับระบบการเรียนรู้การคัดกรองเว็บไซต์ ให้ W คือเวกเตอร์น้ำหนักของการเรียนรู้ ค่า b คือค่าไบอัส (bias) สำหรับระบบการเรียนรู้และ x เป็นเวกเตอร์ข้อมูลที่ใช้สอนในระบบการเรียนรู้การคัดกรองเว็บไซต์ สมการที่เป็นตัวบ่งบอกว่าข้อมูลอยู่ด้านใด คือ

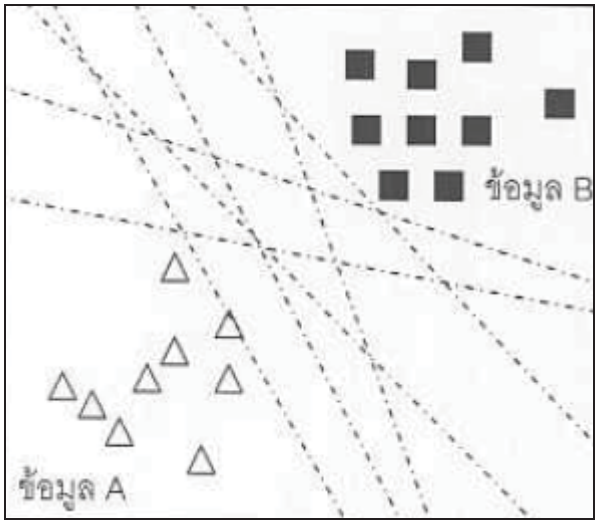
$$f(x) = \text{sign}(wx+b) \tag{2.13}$$

เมื่อเราแทนค่า x ลงในสมการ ถ้า $f(x) = 1$ จะได้คำตอบของชุดข้อมูลอยู่ใน Class 1 ถ้า $f(x) = -1$ จะได้คำตอบของชุดข้อมูลอยู่ใน Class -1

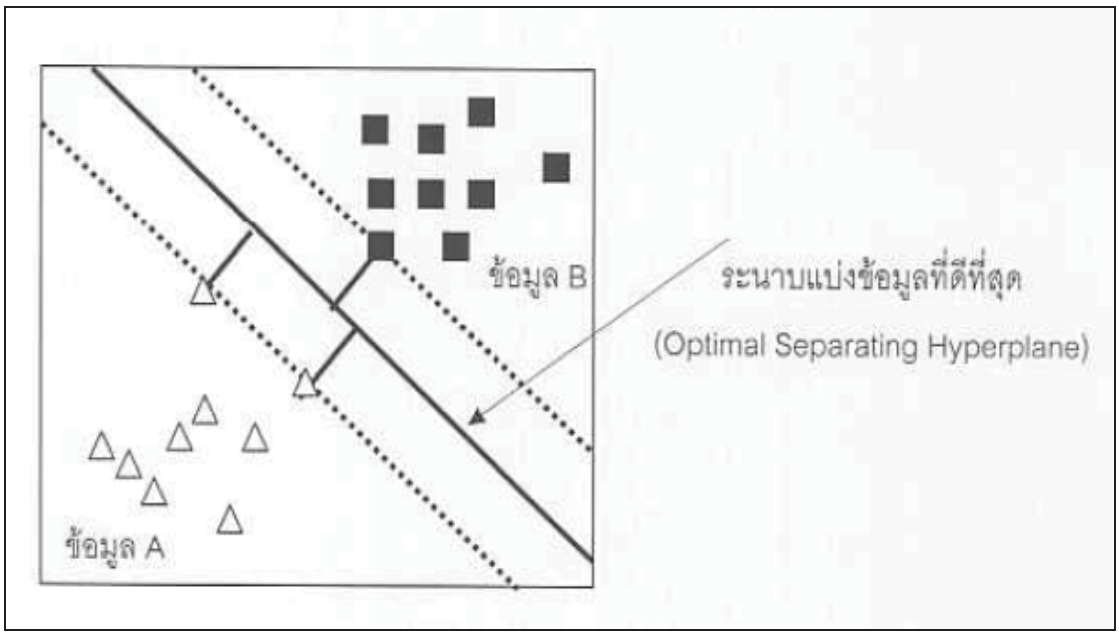


ภาพที่ 2 แสดงสมการเชิงเส้นที่ใช้แบ่งกลุ่มข้อมูล

โดยจุดประสงค์ของ SVM คือการสร้างระนาบที่สามารถแบ่งชุดข้อมูลได้ดีที่สุด (Optimal Separating Hyperplane)



ภาพที่ 3 แสดงเส้นแบ่งข้อมูลในระบบทั่ว ๆ ไป
ที่มา : ธีรพงศ์ โหมดหิรัญ, “การแก้ไขปัญหาความกำกวมของคำในภาษาไทยโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน” (วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, 2548), 37.

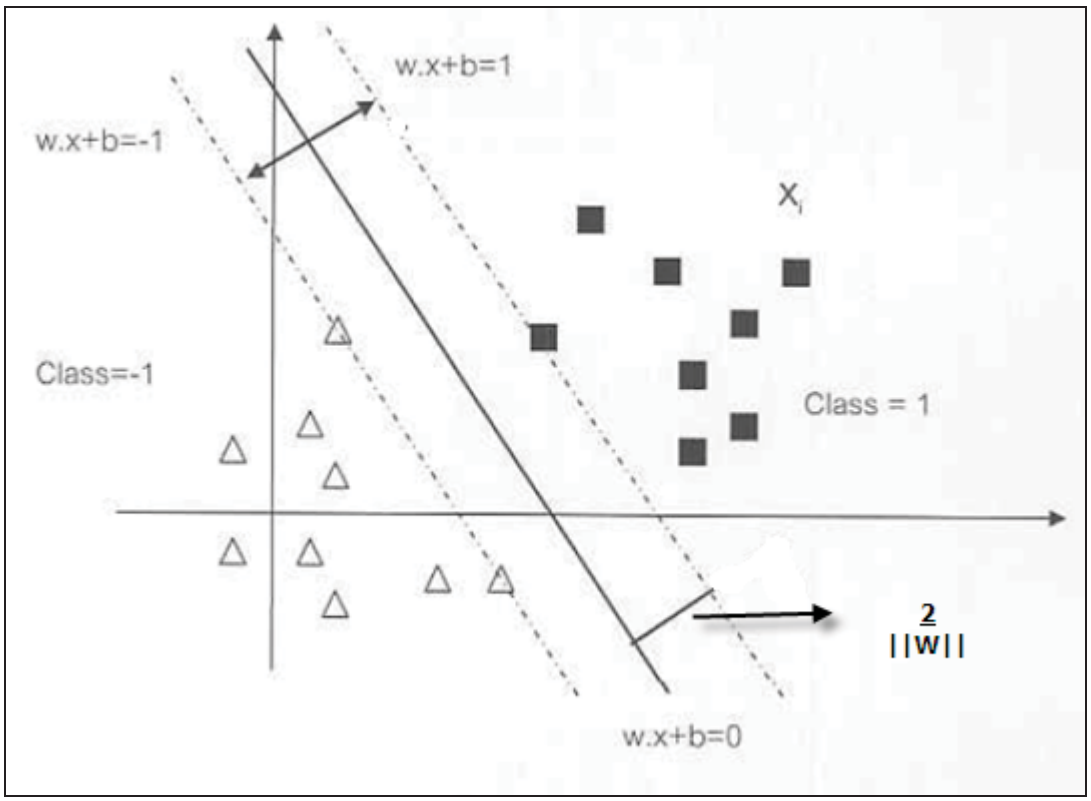


ภาพที่ 4 แสดงเส้นแบ่งข้อมูลของซัพพอร์ตเวกเตอร์แมชชีน
ที่มา : ธีรพงศ์ โหมดหิรัญ, “การแก้ไขปัญหาความกำกวมของคำในภาษาไทยโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน” (วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, 2548), 37.

3.2 การกำหนดค่า Margin Maximization

3.2.1 การหาค่า Margin โดยใช้ทฤษฎีการหาระยะห่างระหว่างจุดกับเส้นตรง

กำหนดให้ลักษณะของข้อมูลที่ใช้ในการเรียนรู้ที่มีขนาด L เป็น (x_i, y_i) ซึ่ง $x_i \in \mathrm{R}^{\mathrm{N}}$ โดยที่ i = 1,…., λ และ $y_i \in \{-1,1\}$



ภาพที่ 5 แสดงค่า Margin ระหว่างจุดข้อมูลใด ๆ กับระนาบ Hyperplane
ที่มา : ธีรพงศ์ โหมดหิรัญ, “การแก้ไขปัญหาความกำกวมของคำในภาษาไทยโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน” (วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, 2548), 38.

ระนาบ wx+b = 0 ประกอบด้วย w ซึ่งกำหนดทิศทาง (Projection) ของระนาบ และ b ซึ่งเป็นค่าเปลี่ยนแปลงบนแกน w โดยระบบจะทำการหาค่ากว้างของเส้นขอบซึ่งเป็นระยะห่างที่ตั้งฉากระหว่างจุดใด ๆ ไปหาสมการเส้นตรง wx + b = 0

โดยให้ระนาบ w.x + b = 0 และกำหนดจุด (x,y) เป็นจุดหนึ่งของข้อมูล

จากทฤษฎีการหาระยะห่างที่ตั้งฉากระหว่างเส้นตรงกับจุด ถ้าให้ d เป็นระยะห่างระหว่างเส้นตรงกับจุดและ Ax + By + C = 0 เป็นสมการเส้นตรง และ (x_1 , y_1) เป็นจุด จะได้ (เรขาคณิตวิเคราะห์, 2553)

$$d = \frac{| Ax_1 + By_1 + C |}{\sqrt{A^2 + B^2}} \qquad (2.14)$$

จากทฤษฎีการหาระยะห่างระหว่างเส้นตรงกับจุดดังนั้นจะได้

$$\frac{|wx_1 + b|}{\sqrt{w.w}}$$

และจากสมการ wx + b = 1 จะได้ $\frac{|1|}{\sqrt{w.w}}$

ในทำนองเดียวกันหาค่า d ภายใต้สมการ

wx + b = -1 จะได้ $\frac{|-1|}{\sqrt{w.w}}$

ดังนั้นจะได้ค่า Margin $= \frac{|1|}{\sqrt{w.w}} + \frac{|-1|}{\sqrt{w.w}} = \frac{|2|}{\sqrt{w.w}}$

3.2.2 การหาค่า Maximization Margin โดยใช้ทฤษฎี Lagrange Multiplier

ทฤษฎีที่ใช้ในการหาค่าเหมาะสมที่สุดเช่น ศึกษาค่า Maximum และ ค่า Minimum ซึ่งไว้คำนวณงานที่มีลักษณะ ตัวแปรตามไม่ได้ขึ้นอยู่กับเพียงตัวแปรต้นเพียงตัวเดียว แต่ขึ้นอยู่กับตัวแปรต้นหลายตัว ซึ่งวิธีการแก้ไขปัญหาจะใช้ ทฤษฎี Lagrange Multiplier ซึ่งต้องทำภายใต้เงื่อนไขข้อจำกัดบางอย่าง (อสมการและเอกลักษณ์ เกี่ยวกับ convexity, 2006)

จากสมการ $wx + b \geq y$ กำหนดให้ $y = 1$ (2.15)

$wx + b \leq y$ กำหนดให้ $y = -1$ (2.16)

จะได้สมการใหม่จากการพิจารณาสมการ (2.15) และสมการ (2.16) ซึ่งระนาบใหม่ที่เกิดขึ้นนี้เป็นระนาบที่ดีที่สุดในการแบ่งข้อมูลออกจากกัน โดยที่ข้อกำหนดในการแบ่งกลุ่มทั้ง 2 กลุ่มออกจากกันโดยที่ปราศจากข้อผิดพลาดคือ

$y_i(wx_i + b) > 0$ สำหรับ $i = 1,\ldots,\lambda$ (2.17)

หรือ

$y_i(wx_i + b) \geq 1$ เมื่อ $\min_{i=1,\ldots,\lambda} |wx_i + b| = 1$ (2.18)

สำหรับทุก ๆ x_i และ $y_i \in \{-1,1\}$ ซึ่งสมการที่ (2.18) จะใช้เป็นเงื่อนไขในทฤษฎี Lagrange Multiplier

และในบรรดาระนาบที่ทำให้เงื่อนไข สมการที่ (2.18) เป็นจริงนั้น ระนาบที่ดีที่สุด คือ ระนาบที่มีระยะห่าง (Margin) ระหว่างกลุ่มทั้ง 2 กลุ่มมากที่สุดโดยที่ระยะห่างนี้ขึ้นอยู่กับค่า 2/||w|| ซึ่งเราสามารถหาคู่ของระนาบซึ่งให้ระยะห่างมากที่สุด (Maximum Margins) โดยการลดค่าของ $\frac{1}{2}||w||$ ให้ต่ำที่สุด ดังนั้นการหาระนาบที่ดีที่สุดทำได้โดยการแก้ปัญหา

$$\text{Minimize } (w,b) = \frac{1}{2}\|w\| \tag{2.19}$$

ภายใต้เงื่อนไข $y(w\,x + b) \geq 1$, $i = 1,\ldots,\lambda$

จาก ทฤษฎี Lagrange Multiplier การหาค่า Max/ Minของสมการ f(x,y,z) ภายใต้เงื่อนไข $g(x,y,z) = k$

จะได้สมการ Lagrange คือ $F(x,y,z,\lambda) = f(x,y,z) - \lambda\,[g(x,y,z) - k]$ (2.20)

ดังนั้นตามทฤษฎี Lagrange จะได้ $L = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i[y_i(w_{.i}\,x + b) - 1]$ (2.21)

เมื่อ α_i คือ Lagrange Multipliers $\alpha_i \geq 0$; i = 1,…..,N

ทฤษฎี Optimization บอกเราว่าค่าระนาบ (w,b) ที่ให้ค่าสมการ (2.21) น้อยที่สุดนั้นจะต้องมีค่าดิฟเฟอร์เรนเชียล (Differential) ของฟังก์ชั่น L ที่จุด w และ b นี้เท่ากับ 0 สมการ (2.21) จึงถูกนำมาหาอนุพันธ์ (Differential)

$$\frac{\partial L(w,b,\alpha)}{\partial w} = w - \sum_{i=1}^{n} y_i\alpha_i x_i = 0 \tag{2.22}$$

และ

$$\frac{\partial L(w,b,\alpha)}{\partial b} = -\sum_{i=1}^{n} y_i\alpha_i = 0 \tag{2.23}$$

ดังนั้นสามารถหาคำตอบของเวกเตอร์ที่ดีที่สุดเพื่อนำไปใช้กับสมการที่ (2.13)

$$w = \sum_{i=1}^{n} y_i\alpha_i x_i \tag{2.24}$$

โดยที่การหาค่า α นั้นเราจะใช้สมการควบคู่ Dual formulation ซึ่งสมการควบคู่นี้ ได้มาจากการนำสมการที่(2.23) และ สมการที่ (2.24) แทนค่าในสมการที่ (2.21)จะได้

$$F(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\|w\|^2 = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\,\alpha_j\,y_i\,y_j\,(x_i\cdot x_j) \tag{2.25}$$

เมื่อ $\alpha_i \geq 0$ $i = 1,\ldots\ldots,l$ และ $\sum_{i=1}^{n} y_i\alpha_i = 0$

สำหรับทฤษฎี Optimization บอกว่า α นั้นได้มาจากการหาค่าสูงสุดของ $F(\alpha)$ โดยที่ค่านี้ต้องอยู่บนเงื่อนไขสมการที่ (2.23)

ในการแก้ไขปัญหาสมการที่ 10 นั้นจะมีค่า Lagrange Multipliers ($(\alpha = (\alpha_1,\ldots.,\alpha_\lambda))$ ที่เป็น 0 อยู่หลายค่า การแก้ไขสมการที่ (2.25) นั้นทำได้โดยการใช้เงื่อนไข KKT (KKT : karush-kuhn-Tucker) Condition ซึ่งจะได้ $y_i(w.\,x_i + b) = 1$ เมื่อค่า $\alpha_i \neq 0$ เมื่อเราได้ค่า α มาแล้วเราสามารถหาค่า b ด้วยการแทนค่า α_i

$$b^* = \frac{1}{\#SV} \sum_{i=1}^{l} (y_i - w^* . x_i) \quad (2.26)$$

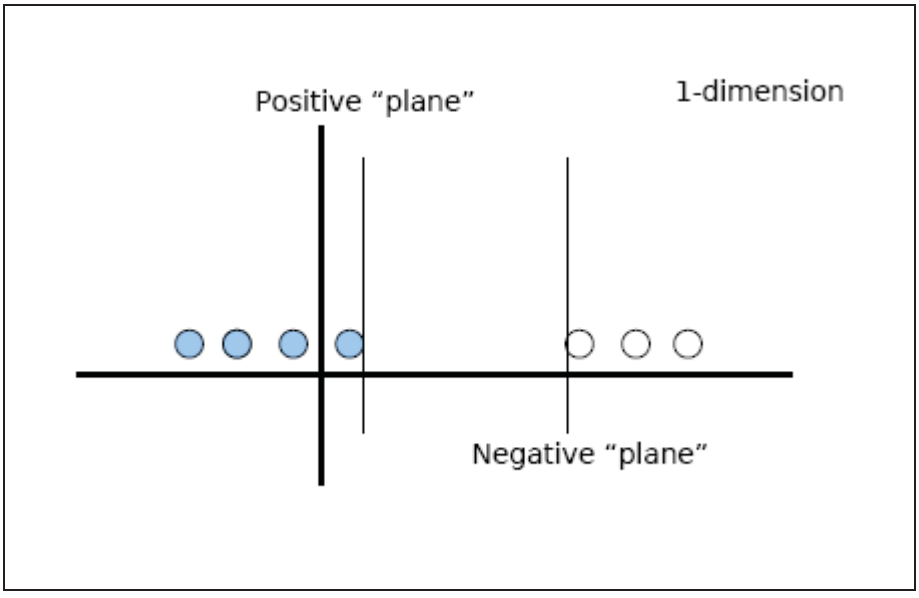
ซึ่งสมการที่ (2.26) เป็นการหาค่าเฉลี่ยของ b^* (เมื่อ #SV = จำนวน support vector) เมื่อแทนสมการที่ (2.24) และ สมการที่ (2.26) จะได้สมการที่ใช้ในการแบ่งกลุ่มข้อมูลเป็น

$$F(x) = \text{sign} (\sum_{i=1}^{l} y_i \alpha_i x x_i + b^*) \quad (2.27)$$

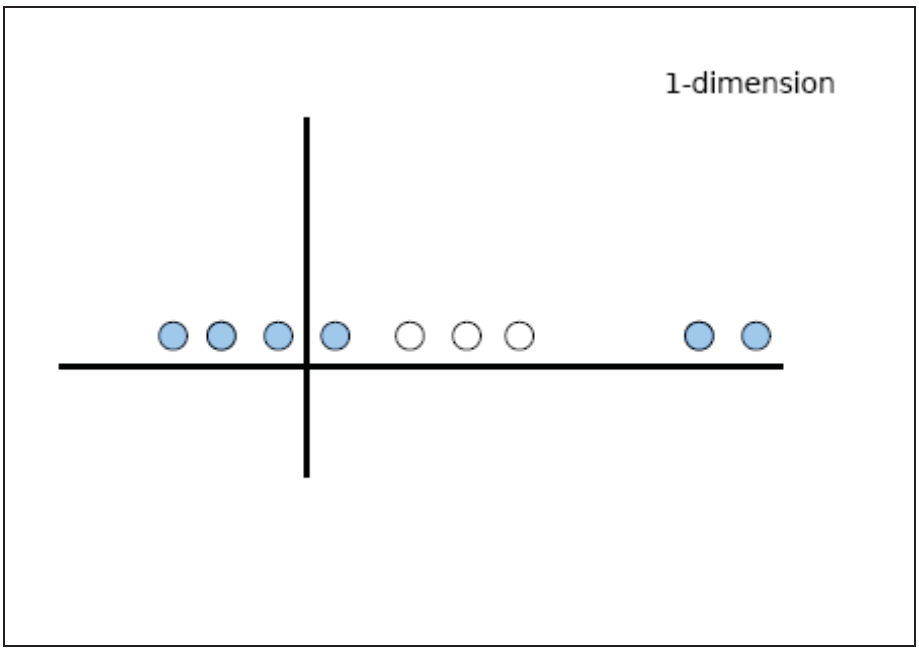
จากสมการที่ (2.27) ค่า y_i คือคำตอบ (Class) ของ Support Vector ที่ x_i และ x คือข้อมูลชุดทดสอบ ค่า α_i และค่า b^* เป็นค่าสัมประสิทธิ์ที่ได้มาจากการคำนวณ (Optimization) ที่กล่าวไปแล้ว

3.3 ทฤษฎี Feature space (อานนท์ นามสนิท 2548 : 41)

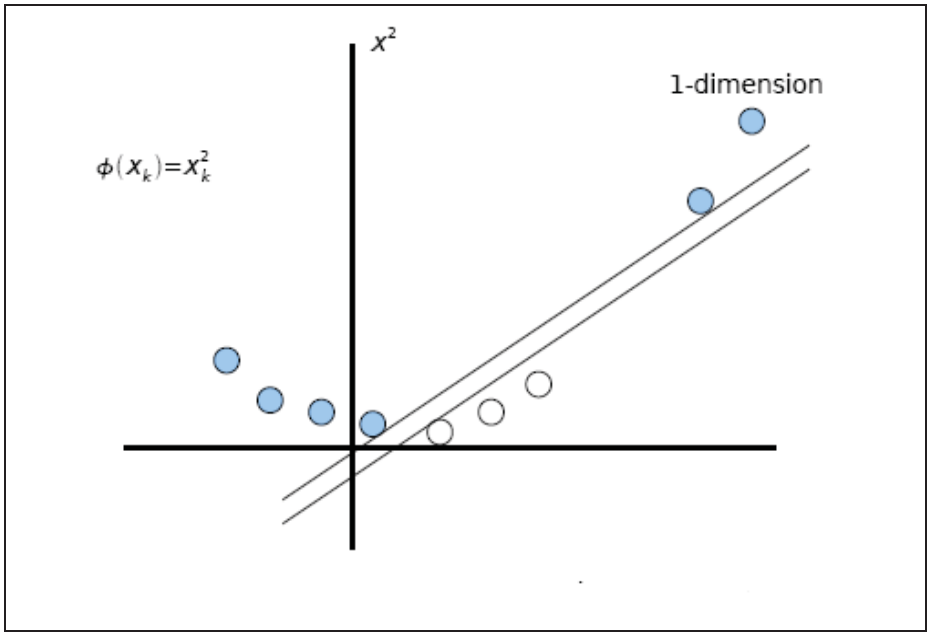
กรณีเว็บไซต์ทั้ง 2 กลุ่มไม่ได้วางตัวใน Feature space ที่สามารถทำการแบ่งได้ด้วยสมการเชิงเส้นแต่ข้อมูลอาจจะจับกลุ่มในตำแหน่งต่าง ๆ ดังนั้นจึงเป็นปัญหาทำให้ไม่สามารถที่จะใช้สมการซัพพอร์ตเวคเตอร์แมชชีนแบบเชิงเส้นได้ จึงต้องมีการเปลี่ยน Feature space เดิมเพื่อให้สามารถแบ่งข้อมูลได้



ภาพที่ 6 กรณีที่ข้อมูลสามารถแบ่งได้ด้วยสมการเชิงเส้น



ภาพที่ 7 กรณีที่ข้อมูลไม่สามารถแบ่งได้ด้วยสมการเชิงเส้น



ภาพที่ 8 การปรับ Feature Space เพื่อให้ข้อมูลสามารถแบ่งได้ด้วยสมการเชิงเส้น

3.4 Kernel Functions

ทฤษฎีที่ทำการ Mapping Feature ของ inner product ในสมการซัพพอร์ตเวคเตอร์แมชชีนแบบเชิงเส้นให้เป็น inner product ใหม่

$$x_i . x_j \rightarrow \phi(x_i).\phi(x_j) \quad (2.28)$$

โดย Function ในการ mapping นั้นจะใช้ทฤษฎี Kernel Functions

$$K(x_i . x_j) = \phi(x_i).\phi(x_j) \quad (2.29)$$

กรณีเปลี่ยน Feature space ใหม่จะได้สมการในการแบ่งกลุ่มใหม่ดังนี้

$$F(x) = \text{sign}\left(\sum_{i=1}^{l} y_i \alpha_i \phi(x)\phi(x_i) + b\right) \quad (2.30)$$

ฟังก์ชั่น Kernel มีดังนี้

1. linear $K(x_i . x_j) = (x_i . x_j)$ (2.31)
2. polynomial $K(x_i . x_j) = K(x_i . x_j)^d$ (2.32)
3. Radial Basis Function (RBF) $K(x_i . x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{g^2}\right)$ (2.33)
4. Sigmoid $K(x_i . x_j) = \tanh(s(x_i . x_j) + c)$ (2.34)

3.5 ขั้นตอนการวัดประสิทธิภาพของระบบ SVM (จันทิมา พลพินิจ และอุมาภรณ์ สายแสงจันทร์ 2548 : 11)

เป็นขั้นตอนของการนำผลเว็บไซต์ที่ทำการแบ่งประเภทเว็บไซต์แล้วมาประเมินประสิทธิภาพ โดยจะตรวจสอบดูว่าเว็บไซต์ที่แบ่งประเภทนั้นมีค่าเป็นอย่างไร โดยทำการประเมินประสิทธิภาพค่าต่าง ๆ ดังนี้

3.5.1 ค่าความถูกต้อง Accuracy คือ ประเภทเว็บไซต์ที่ทำการแบ่งได้นั้นตรงกับประเภทเว็บไซต์ที่กำหนดหรือไม่โดยมีจำนวนเว็บไซต์ที่ประเมินตรงกับที่กำหนดกี่เว็บไซต์

3.5.2 ค่าความแม่นยำ (P) Precision เป็นอัตราส่วนของการค้นพบเว็บไซต์ที่ถูกต้องจากจำนวนเว็บไซต์ทั้งหมดที่ทำการค้นหามาได้

$$\frac{\text{จำนวนเว็บไซต์ที่ถูกต้องและค้นคืนได้}}{\text{จำนวนเว็บไซต์ทั้งหมดที่ค้นคืนได้}}$$

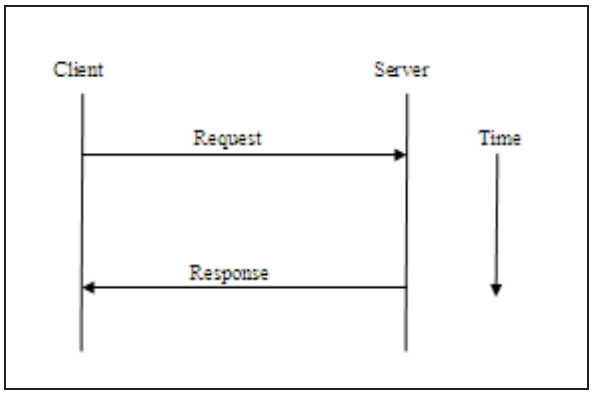
3.5.3 ค่าความระลึก (R) Recall เป็นอัตราส่วนของการค้นพบเว็บไซต์ที่ถูกต้องจากจำนวนเว็บไซต์ที่ถูกต้องทั้งหมด

จำนวนเว็บไซต์ค้นคืนได้

จำนวนเว็บไซต์ทั้งหมดที่ทำการทดสอบ

4. ขั้นตอนการติดต่อกับ Web Server (พิชานี ทัศนเสถียร และภักดิ์ทูล ใจทอง 2549 : 26-28)

รูปแบบที่ใช้ในการสื่อสารระหว่าง Browser กับเว็บไซต์เป็นการสื่อสารในรูปแบบของ Request Response นั่นคือ Browser ซึ่งอยู่ฝั่งของ Client ส่งคำร้องขอหรือ Request ไปยังเว็บไซต์ ซึ่งอยู่ฝั่งของ Server เมื่อ Server ได้รับคำร้องขอแล้ว Server จะส่ง response ซึ่งก็คือคำตอบ พร้อมกับข้อมูลที่ Client ต้องการกลับไปให้โดยมีการทำงานดังภาพที่ 9



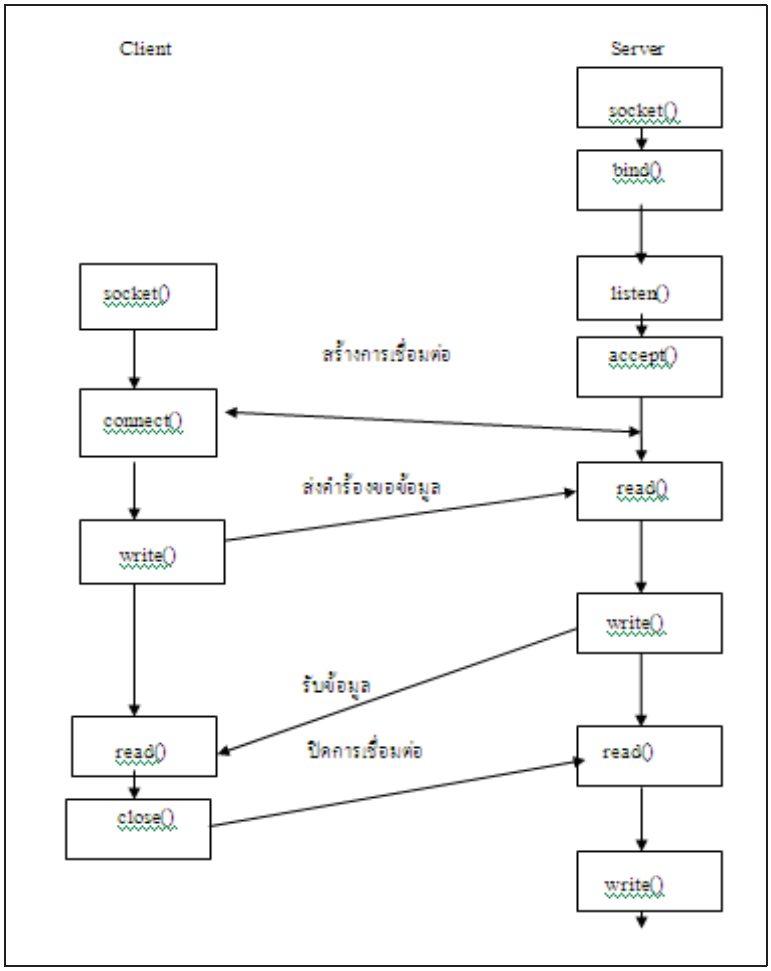
ภาพที่ 9 การส่ง Request ไปยัง Web Server และการรับข้อมูลจาก Web Server
ที่มา : กำธร สุทธิรัตน์, “ระบบป้องกันการเข้าถึงเว็บที่ไม่เหมาะสม : กรณีศึกษาโรงเรียนเทพมงคลรังสี จังหวัดกาญจนบุรี” (สารนิพนธ์ปริญญามหาบัณฑิต สาขาวิทยาการมคอมพิวเตอร์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร, 2546), 13.

การส่ง Request ไปยัง Web server และการรับข้อมูลจาก Web server

การติดต่อสื่อสารระหว่าง Client และ Server เพื่อส่ง Request ไปร้องขอข้อมูลหน้าเว็บเพจของเครื่อง Client ประกอบไปด้วยขั้นตอนดังต่อไปนี้

1. เปิด socket สำหรับติดต่อผ่านระบบ network
2. สร้าง connection กับ Web Server
3. ส่ง Request คำร้องขอไปยัง Web Server
4. รอรับ Response และข้อมูลจาก Web Server
5. ปิด socket ที่ใช้ในการติดต่อ

สรุปเป็นขั้นตอนการติดต่อสื่อสารระหว่าง Client และ Server ในรูปของการใช้คำสั่งของการเขียนโปรแกรมคอมพิวเตอร์เป็นดังภาพที่ 10



ภาพที่ 10 ขั้นตอนการติดต่อสื่อสารระหว่าง Client และ Server
ที่มา : กำธร สุทธิรัตน์, “ระบบป้องกันการเข้าถึงเว็บที่ไม่เหมาะสม : กรณีศึกษาโรงเรียนเทพมงคลรังสี จังหวัดกาญจนบุรี” (สารนิพนธ์ปริญญามหาบัณฑิต สาขาวิทยาการมคอมพิวเตอร์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร, 2546), 14.

HTTP Return Code

ในการติดต่อสื่อสารระหว่าง Client และ Server เพื่อส่ง Request ขอข้อมูล HTTP จะมีรหัสตอบกลับมาจากฝั่ง Server เรียกว่า HTTP Return Code ซึ่งจะมีความหมายต่างๆ กันไปดังนี้

1xx information - การรับคำร้องขอ และรอทำงานต่อไป

2xx successful - ได้รับข้อมูลแล้ว, เข้าใจ และ ยอมรับ

3xx redirection – มีการส่งต่อการทำงานเพื่อให้ได้ข้อมูลตามคำร้องขอ

4xx client error – คำร้องขอผิดรูปแบบ หรือไม่สมบูรณ์

5xx internal server error – เกิดข้อผิดพลาดในเครื่องแม่ข่าย ไม่สามารถปฏิบัติตามคำร้องขอได้

5. การจำแนกประเภทของการกรองเว็บไซต์ที่ไม่เหมาะสม (กำธร สุทธิรัตน์ 2549 : 4)

การกรองเว็บไซต์ที่ไม่เหมาะสมสามารถแบ่งได้เป็น 2 ประเภท ได้แก่

5.1 ระบบการกรองเว็บไซต์ที่ไม่เหมาะสมที่เครื่อง Client

เป็นวิธีการทำงานร่วมกับการทำงานของโปรแกรม Browser โดยการสร้างระบบ Rating ให้กับเว็บเพื่อระบุประเภทของเนื้อหาภายในเว็บและแยกเป็นประเภทเก็บไว้ และกำหนดระดับความสามารถในการใช้งานให้กับผู้ใช้ เพื่อควบคุมขอบเขตการใช้งาน จะมีองค์กรกลางในการรวบรวมรายชื่อเว็บไซต์และกำหนด Rating เพื่อใช้เป็นฐานข้อมูลกลางให้กับผู้ใช้ ปัจจุบันมีหน่วยงานชื่อ Recreational Software Advisory Council's Internet rating system (RSACi) สร้างระบบ Rating ขึ้นมาใช้งานร่วมกับ Internet Explorer ซึ่งเป็น Browser ที่ติดมากับ Windows สามารถกำหนดระดับความสามารถในการใช้งานของผู้ใช้ได้

5.2 สร้างระบบการกรองเว็บไซต์ที่ไม่เหมาะสมที่ Gateway

Gateway เป็นเส้นทางที่ใช้ติดต่อกับเครือข่ายอินเทอร์เน็ต ทั้งที่เป็น Router , Firewall และ Proxy Cache Server จะมี Software ทำงานอยู่บน Hardware ที่ทำหน้าที่เป็น Gateway เหล่านี้ เพื่อกรองเว็บไซต์ที่ไม่เหมาะสม ทำให้ Client ที่อยู่ภายใต้ Gateway ไม่สามารถเข้าชมเว็บไซต์เหล่านี้ได้

การสร้างระบบการกรองเว็บไซต์ไม่เหมาะสมที่เครื่อง Client มีความยืดหยุ่นต่อการใช้งานเพราะผู้ใช้สามารถกำหนดระดับความสามารถของตนเองได้ ตามความต้องการของแต่ละ Client ซึ่งจะทำให้เกิดความหลากหลายต่อการใช้งาน ส่วนการสร้างระบบการกรองเว็บไซต์ไม่เหมาะสมที่ Gateway จะเป็นการกำหนดให้กับระบบภายในทั้งระบบ จะมีผลต่อ Client ที่อยู่ภายในทั้งหมดไม่มีความยืดหยุ่น แต่สะดวกต่อการจัดการเพราะทำครั้งเดียว แห่งเดียว แต่มีผลกับระบบทั้งหมด

6. วิธีป้องกันเว็บไซต์ที่ไม่เหมาะสมที่มีอยู่ในปัจจุบัน (Rongbo, Reihaneh and Willy 2003 : 325 – 326)

วิธีป้องกันของระบบการกรองเว็บไซต์ที่ไม่เหมาะสมทั้งแบบสร้างที่เครื่อง Client หรือสร้างที่ Gateway จะมีวิธีป้องกันอยู่ 3 วิธี ด้วยกัน

6.1 Blacklists and Whitelists

วิธีนี้เป็นการสร้างรายชื่อเว็บไซต์ที่ไม่อนุญาตการใช้งานและอนุญาตการใช้งานให้กับระบบ โดยBlacklists จะเป็นรายชื่อเว็บไซต์ที่ไม่อนุญาตให้ใช้และ Whitelists จะเป็นรายชื่อเว็บไซต์ที่อนุญาตให้ใช้งานได้ (Oskar 2003 : 23-26) โดยรายชื่อเหล่านี้จะต้องทำการปรับปรุงอยู่ตลอดเวลา เนื่องจากมีเว็บไซต์ที่เกิดใหม่อยู่เสมอ และเป็นภาระของบุคลากรที่จะต้องคอยปรับปรุง Blacklistsและ Whitelists อยู่ตลอดเวลาเช่นกัน อีกทั้งหากระบบที่อยู่เหนือขึ้นไปมีการกรองเว็บไซต์ ไม่เหมาะสมอยู่แล้วบางส่วน ก็จะทำให้มีการทำงานซ้ำซ้อน เป็นการเพิ่มภาระให้กับระบบมากขึ้น

6.2 Keyboard Blocking

วิธีนี้เป็นการใช้กลุ่มของข้อความต้องห้าม มาเป็นเครื่องมือในการสร้างกฎเกณฑ์ในการอนุญาต หรือไม่อนุญาตให้ใช้ (Oskar 2003 : 23-26) โดยจะใช้กลุ่มของข้อความต้องห้ามเป็นข้อมูล สำหรับเปรียบเทียบกับชื่อเว็บไซต์ ถ้ามีข้อความที่ตรงกับกลุ่มของข้อความต้องห้ามอาจจะตรงกันบางส่วนหรือตรงกันทั้งหมด ก็จะไม่อนุญาตให้ใช้งานในเว็บไซต์ดังกล่าว แต่ถ้าไม่ตรงกับกลุ่มของข้อความต้องห้าม ก็จะอนุญาตให้ใช้งานเว็บไซต์นั้นได้ วิธีการนี้ถ้าผู้สร้างเว็บไซต์ไม่เหมาะสมอาศัยการตั้งชื่อแบบสะกดผิด อาจหลีกเลี่ยงวิธีการกรองแบบนี้ได้ และการเลือกกลุ่มของข้อความที่จะนำมาใช้สร้างกฎเกณฑ์ในการอนุญาตหรือไม่อนุญาต ก็จะเป็นการยากที่จะเลือกให้ได้ครอบคลุม เพราะคำบางคำอาจมีความหมายทั้งสองด้าน

6.3 Rating System

วิธีนี้เป็นการกำหนด Rating ให้กับเว็บไซต์ต่างๆ โดยผู้ดูแลเว็บจะพิจารณากำหนด Rating ของเว็บไซต์ของตน และนำชื่อเว็บไซต์ของตนไปลงทะเบียนการจัด Rating ไว้กับองค์กรกลางที่รับจดทะเบียน เช่น The Recreational Software Advisory Council ratingservice for Internet (RSACi) และ Internet Content Rating Association (ICRA) องค์กรกลางเหล่านี้จะเป็นศูนย์กลางในการรวบรวมรายชื่อเว็บไซต์ และจัด Rating ตามที่ได้ลงทะเบียนเอาไว้ การใช้งานจะทำงานผ่านระบบ Browser หรือ Application ที่ทำงานร่วมกับ Browser (พิรงรอง รามสูตรณะนันทน์และนิธิมา คณานิธินันท์ 2547 : 33-35)

7. ระบบการป้องกันเว็บไซต์ที่ไม่เหมาะสมที่มีในปัจจุบัน

ปัจจุบันมีหลายหน่วยงานที่พัฒนาการป้องกันเว็บไซต์ที่ไม่เหมาะสม ได้แก่

7.1 The Internet Content Rating Association (ICRA)

เป็นองค์กรที่ไม่หวังผลกำไร ริเริ่มจาก Dr. Donald F. Roberts จากมหาวิทยาลัย Standford สร้างระบบการป้องกันเว็บไซต์ที่ไม่เหมะสมด้วยวิธี Rating System ใช้งานผ่าน Browser โดยมี ICRAplus filter เป็นเครื่องมือในการกรองเว็บไซต์ไม่เหมาะสม เพื่อป้องกันเด็กและเยาวชนจากพิษภัยที่เกิดจากอินเทอร์เน็ต เดิมใช้ชื่อว่า Recreational Software Advisory Council's Internet rating System (RSACi) (w.w.w.rsac.org) (Family Online Safety Institute 2006)

7.2 SafeSurf Rating

เป็นระบบป้องกันเว็บไซต์ไม่เหมาะสมที่สร้างด้วยวิธี Rating System ใช้งานผ่าน Browser ที่แบ่งเนื้อหาของเว็บไซต์ออกเป็น 12 ด้าน เป้าหมายหลักเพื่อให้มีการใช้งานอินเทอร์เน็ตได้อย่างเหมาะสม ต่อมาได้พัฒนาเพื่อสนับสนุนช่วยเหลือและให้ความเป็นธรรมกับผู้ปกครองของเด็กและเยาวชน (SafeSurf 2006)

7.3 SmartFilter

เป็นระบบป้องกันเว็บไซต์ไม่เหมาะสมที่สร้างขึ้นสำหรับอุปกรณ์ที่ทำหน้าที่เป็น Firewall, Proxy หรือ Caching มีลักษณะเป็นสินค้าที่เป็นอุปกรณ์สำหรับใช้งานโดยมี Software บรรจุอยู่ในตัวอุปกรณ์ เน้นการป้องกันการใช้เครือข่ายไปในทางที่ไม่ถูกต้อง (Secure Computing Corporation 2006)

7.4 Squidguard

เป็น Software ที่ติดตั้งเพิ่มเติมเพื่อใช้งานกับ Proxy Cache Server ใช้งานควบคู่ไปกับ SQUID สามารถจำกัดการใช้งาน อนุญาตให้ใช้งานเฉพาะเว็บไซต์ที่เหมาะสม ห้ามใช้งานเว็บไซต์ที่ไม่เหมาะสม (blacklists) ห้ามใช้งานเว็บไซต์ที่มีคำต้องห้ามอยู่ในชื่อของเว็บไซต์ สามารถ update Blacklist กับทางเว็บไซต์ได้ (SquidGuard 2006)

7.5 Internet Access Content Management

เป็นอุปกรณ์บริหารจัดการการใช้อินเทอร์เน็ตภายในองค์กร สามารถป้องกันการเข้าไปดูเว็บไซต์ไม่เหมะสมได้ 60 ประเภท และป้องกันการดาวน์โหลดไฟล์หนังหรือเพลงต่างๆ ที่ทำให้ระบบเครือข่ายทำงานช้าลง และสามารถตรวจสอบการใช้อินเทอร์เน็ตของบุคลากรได้แบบ Real – Time (บุญสูงเทคโนโลยี 2549)

7.6 DataReactor iMimic Networking Inc.

เป็นระบบที่ทำงานบน FreeBSD, Linux และ Solaris ทำงานร่วมกับระบบ Caching ช่วยในการลดภาระงานของ Server และลดปริมาณ Traffic ในระบบ (iMimic Networking 2006)

7.7 SITA URL filtering

เป็นระบบ Web Filtering ติดตั้งพร้อมกับอุปกรณ์ firewall ใช้งานในหน่วยงาน ที่มีความต้องการบังคับ ให้การใช้งานทรัพยากรเครือข่ายที่มีการใช้งานร่วมกันใช้ได้อย่างคุ้มค่า กำหนดกฎเกณฑ์การใช้งานให้กับพนักงานได้เหมาะสมกับงาน สามารถ update URL list ให้ทุกวัน ลักษณะการทำงานเป็นแบบควบคุมการใช้งานจาก Blacklists (SITA 2005)

7.8 WEB Filtering for WinProxy

เป็นระบบที่ทำงานกับ WinProxy ซึ่งเป็นโปรแกรมประเภท Caching Server ทำงานบน Windows 98, NT, 2000, ME, XP มีการ update รายชื่อเว็บไซต์ไม่เหมาะสมอย่างต่อเนื่อง เพื่อป้องกันผู้ใช้ที่ดีออกจากสิ่งที่ไม่ดีในระบบเครือข่าย ลักษณะการทำงานเป็นแบบควบคุมการใช้งานจาก Blacklists (Blue Coat Systems 2006)

7.9 SonicWALL Content Filtering Service (CFS)

เป็นระบบที่ที่มุ่งหวังให้การใช้งานเครือข่ายอินเทอร์เน็ตเป็นไปเพื่อองค์กรและการศึกษาโดยการควบคุมการใช้งานผ่านระบบ Blacklists โดยมีการ update รายชื่อให้อัตโนมัติ (SonicWALL 2006)

7.10 FORTIGUARD WEB Filtering

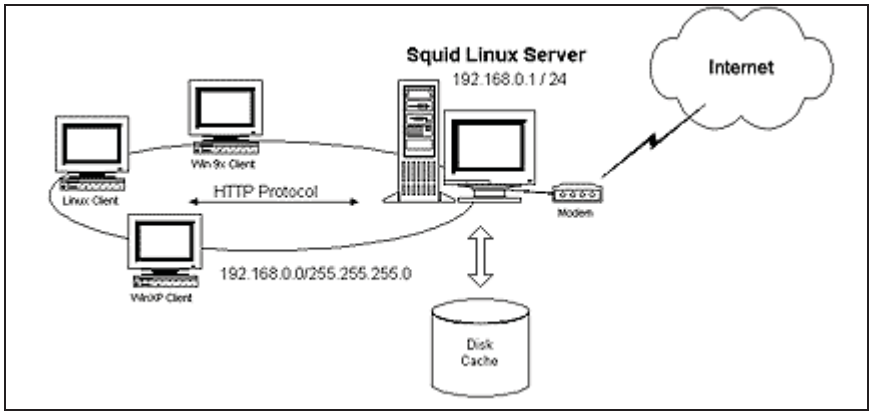
เป็นระบบที่ประกอบด้วย FortiGuard Rating Server เก็บรายชื่อเว็บไซต์ต่างๆไว้พร้อมการกำหนด Rating ในแต่ละเว็บไซต์ และ FortiGuard Antivirus Firewall เป็นส่วนที่ควบคุมการใช้งานเครือข่ายอินเทอร์เน็ตขององค์กร (FORTINET 2006)

นอกจากนี้ยังมีเครื่องมือที่ใช้ในการป้องกันเว็บไม่เหมาะสมอีกหลายตัวด้วยกัน เช่น BizGuard, Cyber Patrol www.cyberpatrol.com, CYBER sitter www.cybersitter.com, Cyber Snoop, Internet Watcher 2000, Net Nanny www.netnanny.com, Norton Internet Security, Optenet, SurfMonkey, X-Stop โดยทั้งหมดมุ่งเน้นไปทางด้านธุรกิจ มีค่าใช้จ่ายในการใช้งานทั้งสิ้น

8. ระบบพร๊อกซี เซิร์ฟเวอร์(Proxy server) (ธีรภัทร มนตรีศาสตร์ 2552)

พร๊อกซี เซิร์ฟเวอร์ หรือเรียกว่า แคช (Cache) คือการนำเครื่องคอมพิวเตอร์ที่ให้บริการแก่กลุ่มผู้ใช้ที่อยู่ในบริเวณเดียวกัน และกำหนดให้ผู้ใช้ทุกคนเรียกข้อมูลเว็บ ผ่านเครื่องคอมพิวเตอร์นี้ โดยเครื่องดังกล่าวจะมีการติดตั้งโปรแกรมเพื่อทำหน้าที่เรียกข้อมูลเว็บมาให้บริการแก่ผู้ใช้และจัดเก็บข้อมูลที่เคยถูกเรียกนั้นไว้ในเครื่อง เพื่อให้บริการข้อมูลแก่ผู้ใช้ข้อมูลนั้นซ้ำได้ทันทีโดยไม่ต้องเสียเวลาไปเรียกข้อมูลมาจากแหล่งข้อมูลใหม่ ซึ่งวิธีนี้ทำให้ผู้ใช้สามารถเรียกใช้ข้อมูล

ที่เคยมีผู้เรียกใช้มาก่อนได้รวดเร็วขึ้น ทำให้ประสิทธิภาพในการใช้งานระบบเครือข่ายอินเตอร์เน็ตเพิ่มขึ้น



ภาพที่ 11 การทำงานของ Proxy Server

ที่มา : ธีรภัทร มนตรีศาสตร์, Squid Proxy Caching Server [ออนไลน์], เข้าถึงเมื่อ 8 พฤศจิกายน 2551. เข้าถึงได้จาก http://micro.se-ed.com/content/mc205/MC205_181.asp

ระบบพร๊อกซี เซิร์ฟเวอร์มีโปรแกรมที่นิยมใช้มากที่สุด และมีความสามารถสูงคือ Squid (www.squid-cache.org) ซึ่งจะมีมาพร้อมกับ Linux Server ทุกตัว โปรแกรม Squid เป็นพร๊อกซีเซิร์ฟเวอร์ ที่มีคุณสมบัติในการจัดการจำกัด ควบคุมการเข้าสู่เว็บไซต์ภายนอกองค์กร ได้เป็นอย่างดี โดยมี Access Control List (ACL) ทำหน้าที่เป็นเสมือนกฎในการเข้าใช้งาน และ Squid ยังมีคุณสมบัติเป็น HTTP Object cache ที่ช่วยเก็บข้อมูลจากเว็บไซต์ภายนอกไว้ในหน่วยความจำของตัวเซิร์ฟเวอร์เองอีกด้วย ช่วยให้การเรียกเว็บไซต์ที่เคยเข้าถึงมาก่อนทำได้รวดเร็วยิ่งขึ้น คุณสมบัติของโปรแกรม Squid ที่เกี่ยวข้องกับงานวิจัย

8.1 การ Block เว็บไซต์โดยการสร้าง Blacklist

สำหรับการ Block เว็บไซต์ด้วยโปรแกรม Squid โดยใช้ลักษณะการสร้างรายชื่อเว็บไซต์ต้องห้ามเอาไว้ สามารถทำได้โดยพิมพ์รายชื่อต้องห้ามเอาไว้ในไฟล์ที่มีลักษณะเป็นเท็กซ์ไฟล์แล้วกำหนดคำสั่งการ Block ไว้ในไฟล์ /ect/squid/squid.conf โดยใช้แทก (TAG) acl และ HTTP_ access ซึ่งมีรูปแบบการใช้งานดังนี้

TAG : acl

acl aclname url_regex [-i] “file” … #regex matching on whole URL

acl (Access List) เป็นการกำหนดบัญชีรายชื่อการใช้งานเพื่อควบคุมการใช้งานต่างๆ ตามความเหมาะสม มีการกำหนดได้หลายประเภทเป็น url_regex (Uniform Resource Locator Regular expressions) หมายถึงชนิดของ acl ที่เป็นบัญชีรายชื่อของที่อยู่ของเว็บไซต์

TAG : HTTP_ access

HTTP_access allow[deny [!]aclname

HTTP_access (Access to the HTTP port) เป็นการกำหนดการอนุญาต (allow) หรือไม่อนุญาต (deny) ให้ใช้งานพอร์ต HTTP ซึ่งก็คือพอร์ตที่ใช้งานติดต่อกับ Server ของเว็บไซต์

8.2 การ Block เว็บไซต์โดยใช้ Keyword

เป็นการ Block เว็บไซต์ ที่มีชื่อบางส่วนเป็น keyword ที่กำหนดไว้ โดยใช้ TAG : acl และ TAG : HTTP_ access เช่นเดียวกับการ block เว็บไซต์โดยการสร้าง Blacklists ต่างกันตรงที่การกำหนด acl ใช้วิธีกำหนด Keyword ลงไปแทนชื่อไฟล์ มีรูปแบบการใช้งาน ดังนี้

acl aclname url_regex [i] Keyword, Keyword2, …

สามารถกำหนด Keyword ได้หลายค่า สำหรับ HTTP_access ใช้งานเหมือนกับการ Block เว็บไซต์โดยการสร้าง Blacklist

8.3 การทำ Transparency เพื่อให้ Client วิ่งเข้าใช้ Proxy อัตโนมัติ

การทำ Transparency คือ การสร้างระบบตรวจจับการใช้งานพอร์ต HTTP ซึ่งใช้พอร์ตหมายเลข 80 บนโปรโตคอล TCP/IP ให้เปลี่ยนทิศทางไปใช้งานพอร์ตที่เป็น Proxy ซึ่งใช้พอร์ต 8080 โดยอัตโนมัติ โดยที่เครื่อง Client ไม่จำเป็นต้องมีการกำหนดค่าการใช้งาน Proxyและไม่สามารถหลีกเลี่ยงการใช้งานโดยไม่ใช้ Proxy ได้

การสร้างระบบ Transparency มีแนวทางในการสร้าง 3 แนวทางด้วยกัน คือ แนวทางแรกกำหนดเงื่อนไขที่ router แนวทางที่สองใช้ smart switching และแนวทางที่สามกำหนดค่าในเครื่องที่ใช้ Squid ทำ Proxy วางไว้ในตำแหน่งที่เป็น Gateway ของระบบแนวทางแรกและแนวทางที่สองนั้นต้องอาศัยอุปกรณ์ที่มีราคาแพง ส่วนแนวทางที่สาม เพียงแค่ใช้วิธีการกำหนดค่าบางอย่างในเครื่องที่ใช้ Squid ทำหน้าที่เป็น Proxy และต้องวางขวางเส้นทางออกอินเทอร์เน็ตของเครื่อง Client โดยต้องใช้ส่วนประกอบในการกำหนดค่าสองส่วนด้วยกันคือ

ส่วนที่ 1 ใช้ iptables ทำหน้าที่เปลี่ยนทิศทางของ package ที่ใช้งานพอร์ต HTTP ให้ไปใช้พอร์ตของ Proxy แทนโดยใช้คำสั่ง

iptables -t nat -A PREROUTING -i eth1 -p tcp --dport 80 -j REDIRECT --to-port 8080

ซึ่งเป็นคำสั่งที่ทำให้มีการเปลี่ยนทิศทาง package ที่มีต้นทางมาจากเครื่อง Client และมีปลายทางไปที่พอร์ต HTTP(80) ให้ไปที่พอร์ต (8080)

ส่วนที่ 2 ใช้ Squid ซึ่งทำหน้าที่เป็น Proxy รับ package ที่ถูกเปลี่ยนทิศทางมาเพื่อให้บริการ ซึ่งต้องกำหนดค่าใน Squid.conf เพื่อทำงานในโหมด Transparency ดังนี้

http_port 8080 transparent

8.4 ไฟล์บันทึกการใช้งาน access.log

ประวัติการใช้งานของเครื่อง Client จะถูกบันทึกไว้ที่ไฟล์ access.log อยู่ที่ /var/log/squid มีข้อมูลทั้งหมด 10 ฟิลด์ แต่ละฟิลด์ประกอบด้วยข้อมูลดังนี้ 2010-01-19 10:24:59 366 127.0.0.1 TCP_MISS/200 8584 GET http://www.google.co.th/ - DIRECT/64.233.181.147 text/html

Timestamp	เวลาที่การติดต่อสิ้นสุดลง ในรูปแบบของ UNIX
Elapsed Time	ช่วงเวลาที่ใช้ในการติดต่อ
Client Address	หมายเลข IP address ของเครื่อง Client
Log Tag/HTTP Code	รายละเอียดของผลของคำร้องขอ/รหัส HTTP ที่ตอบกลับ
Size	ขนาดของข้อมูลที่ส่งให้ Client
Request Method	คำสั่งของคำร้องขอ
URL	ที่อยู่ของเว็บไซต์ที่ร้องขอข้อมูล
Ident	ชื่อผู้ใช้ที่ได้รับการอนุญาตให้ใช้งานผ่าน Client
Hierarchy Data/ Hostname	ได้ข้อมูลตามคำร้องขอมาอย่างไร / จากที่ไหน
Content Type	ชนิดของข้อมูลที่ร้องขอ

9. อัลกอริทึมสำหรับการกรองเว็บไซต์ไม่เหมาะสม

การจัดแบ่งประเภทของกลุ่มเว็บไซต์ประกอบด้วยกลไกในการจัดแบ่งโดยใช้ข้อความกับกลไกในการจัดแบ่งโดยใช้รูปภาพ ในงานวิจัยนี้จะใช้เฉพาะกลไกในการจัดแบ่งโดยใช้ข้อความเนื่องจากเป็นวิธีที่ง่ายไม่ซับซ้อน จากการค้นคว้าพบว่ามีผู้ได้คิดค้นกลไกในการจัดแบ่งประเภทของกลุ่มเว็บไซต์โดยการใช้ข้อความ (Text Classification) ด้วยกันหลายราย ดังมีรายละเอียดดังนี้

9.1 Naive Bayes

เป็นวิธีที่ง่าย มีประสิทธิภาพ จึงมีการใช้กันอย่างแพร่หลาย โดยวิธีนี้จะใช้อัตราความถี่ของความสัมพันธ์ระหว่างคำในข้อความของเว็บไซต์กับกลุ่มคำตัวอย่าง เพื่อแบ่งประเภทของเว็บไซต์ โดยใช้ค่าความแตกต่างของความน่าจะเป็นมาเป็นเครื่องแบ่งประเภท (Yirong and Jing 2005)

9.2 K-Nearest Neighbor

เป็นวิธีที่ใช้การเลือกกลุ่มเว็บไซต์ที่มีความคล้ายกันเพื่อเป็นกลุ่มตัวอย่างในการจัดแบ่งประเภทกลุ่มเว็บไซต์ โดยเว็บไซต์ที่ต้องการตรวจสอบมีความเหมือนกับกลุ่มเว็บไซต์ตัวอย่าง โดยใช้ vector ของกลุ่มคำในเว็บไซต์เป็นตัวแทนในการเปรียบเทียบ (David and Marc 2008)

9.3 Decision Tree

เป็นวิธีที่ใช้ต้นไม้ (tree) เป็นเครื่องมือในการจัดการแบ่งประเภทของเว็บไซต์โดยเริ่มต้นทำการทดสอบความสัมพันธ์ที่เกี่ยวข้องกันของกลุ่มข้อมูลตัวอย่าง เพื่อตัดสินใจในการท่องไปในกิ่งของต้นไม้แต่ละด้าน (node) จนกว่าจะถึงใบของต้นไม้ (leaf) จึงจะได้ประเภทของเว็บไซต์ (Rajeev and Kyuseok 2005)

9.4 Support Vector Machines

ใช้การตัดสินใจโดยใช้ลักษณะภายนอกของเว็บไซต์ โดยแบ่งเป้าหมายหลักของเว็บไซต์ออกเป็นกลุ่ม ๆ และได้ถูกประยุกต์มาใช้ในการจัดแบ่งประเภทของเว็บไซต์โดยใช้การเปรียบเทียบกับ vectors ซึ่งเป็นตัวแทนของเว็บไซต์กับกลุ่มของเว็บไซต์ตัวอย่าง เพื่อแบ่งแยกความแตกต่างกันของเว็บไซต์ ว่าอยู่ในกลุ่มของเว็บไซต์ประเภทใด (Support Vector Machine 2010)

9.5 Classification of hypertext data

ใช้ชุดคำสั่งของ hyperlink, content of linked และ meta data ที่มีอยู่ในเว็บไซต์มาใช้ในการจัดการแบ่งประเภทของเว็บไซต์ ซึ่งให้ความถูกต้องและแม่นยำในการจัดแบ่งประเภทของเว็บไซต์ (Rayid , sean and Yiming 2006)

9.6 Text Classification for hypertext filtering

เป็นการนำ artificial neural network มาใช้ในการกรองหน้าเว็บไซต์ที่มีภาพโป๊เปลือยโดยใช้การรวบรวมหน้าเว็บไซต์ที่มีภาพโป๊เปลือยและไม่มีภาพโป๊เปลือยมาเป็นเครื่องมือในการ Train ระบบ artificial neural network เพื่อจัดการแบ่งประเภทของเว็บไซต์ วิธีการนี้จำเป็นต้องใช้เครื่องคอมพิวเตอร์ที่มีประสิทธิภาพสูงและไม่สามารถสร้างระบบที่ทำงานแบบ real-time ได้ (Pui, Siu and Alvis 2006)

9.7 Web Filtering Using Text Classification

ใช้Vector ซึ่งเป็นตัวแทนของคำที่ปรากฏในเว็บไซต์ มาเปรียบเทียบหาค่าสัมประสิทธิ์ความเหมือนกับกลุ่มตัวอย่างของเว็บไซต์ ถ้ามีค่าสัมประสิทธิ์ความเหมือนสูงกว่า 50 % จะถือว่าเป็นประเภทเดียวกับกลุ่มตัวอย่างของเว็บไซต์ที่ใช้เปรียบเทียบ (Reihaneh, Safavi - Naini and Susilo 2003 : 327-328)

บทที่ 3

วิธีดำเนินการวิจัย

ขั้นตอนการดำเนินการวิจัย สามารถแบ่งได้เป็น 5 ขั้นตอนหลัก คือ

1. ศึกษาวิธีการเขียนโปรแกรมภาษาซีบน Linux Ubuntu 9.04 เพื่อติดต่อสื่อสารขอข้อมูลหน้าเว็บเพจ จากเครื่องแม่ข่ายของเว็บไซต์ได้

2. ศึกษาคุณลักษณะของเว็บไซต์ไม่เหมาะสมเพื่อสร้างกลุ่มคำที่จะนำมาสร้างระบบการกรองเว็บไม่เหมาะสม

3. ศึกษาทฤษฎี PCA (PCA : Principal Component Analysis) เพื่อใช้ในการแบ่งกลุ่มเว็บไซต์

4. สร้างระบบการกรองเว็บไม่เหมาะสมสำหรับเครื่องแม่ข่าย Proxy Cache Server

5. ทดสอบประสิทธิภาพของระบบ

กลุ่มตัวอย่างที่ใช้ในการวิจัย

รายชื่อเว็บไซต์ไม่เหมาะสมประเภทความรุนแรง ยาเสพติดจำนวน 200 เว็บไซต์ (ภาคผนวก ค หน้า 126) รายชื่อเว็บไซต์ไม่เหมาะสมประเภทลามกอนาจารจำนวน 150 เว็บไซต์ (ภาคผนวก ง หน้า 134) และรายชื่อเว็บไซต์ปกติ 200 เว็บไซต์ (ภาคผนวก จ หน้า 141) ซึ่งได้จากการบันทึกการใช้งานของเครื่องลูกข่าย ที่เข้าใช้ Proxy Cache Server จากไฟล์ access.log ของเครื่องแม่ข่ายอินเทอร์เน็ตของโรงเรียนศรีวิชัยวิทยา

เครื่องมือและอุปกรณ์

1. ระบบเครือข่ายอินเทอร์เน็ตโรงเรียนศรีวิชัยวิทยา
2. เครื่องแม่ข่ายอินเทอร์เน็ตที่ทำหน้าที่ Proxy Cache Server มีรายละเอียด ดังนี้

2.1 CPU	Xeon Processor 3G
2.2 RAM	4 GB
2.3 HardDisk	160 GB
2.4 Monitor	17”

2.5 CDROM 52X

2.6 Network Card 10/100/1000 2 ใบ

2.7 Operating System Linux Ubuntu 9.04 , Squid2.7 STABLE

3. เครื่องลูกข่ายอินเทอร์เน็ต ใช้ทดสอบการทำงาน มีรายละเอียด ดังนี้

3.1 CPU Pentium 4 1.0 GHz.

3.2 RAM 512 GB

3.3 HardDisk 40 GB

3.4 Monitor 14”

3.5 CDROM 52X

3.6 Network Card 10/100

3.7 Operating System Microsoft Windows XP Service Pack 3

4. อุปกรณ์เชื่อมเครือข่าย (Switching) พร้อมสายเชื่อมต่อ

โปรแกรมที่ใช้ในงานวิจัย

1. ระบบปฏิบัติการ Linux Ubuntu9.04
2. ระบบปฏิบัติการ Microsoft Windows XP Service Pack 3
3. โปรแกรม VI สำหรับสร้าง/แก้ไข ซอร์สโค้ด
4. โปรแกรม PICO สำหรับสร้าง/แก้ไข ซอร์สโค้ด
5. โปรแกรม GNU C/C++ สำหรับคอมไพล์ภาษาซี
6. โปรแกรม Squid2.7 สำหรับทำหน้าที่เป็น Proxy Cache Server
7. โปรแกรม Matlab R2007a
8. โปรแกรม SVM^{light} Version 6.02

ขั้นตอนการศึกษาวิธีการเขียนโปรแกรมภาษาซีเพื่อติดต่อกับ Web Server

1. ศึกษาวิธีการเขียนโปรแกรมภาษาซีบน Linux Ubuntu9.04
2. ศึกษาวิธีการเขียนโปรแกรมภาษาซีเพื่อติดต่อกับ Web Server
3. ศึกษาวิธีการเขียนโปรแกรมภาษาซีเพื่อขอข้อมูลหน้าเว็บเพจจากเว็บไซต์เฉพาะ HTML Code บันทึกลงไฟล์

ศึกษาคุณลักษณะของเว็บไซต์ที่ไม่เหมาะสม

1. เขียนโปรแกรมภาษาซี เพื่อขอข้อมูลหน้าเว็บไซต์ของเว็บไซต์ที่ไม่เหมาะสม กลุ่มตัวอย่างที่มีในระบบเดิม บันทึก HTML Code ลงไฟล์

2. ศึกษาคุณลักษณะเฉพาะของเว็บไซต์ไม่เหมาะสม สร้างกลุ่มคำตัวอย่างเพื่อนำไปใช้ในการตรวจสอบเว็บไซต์ไม่เหมาะสมต่อไป

อัลกอริทึมสำหรับการสร้างระบบตรวจสอบเว็บไซต์ไม่เหมาะสม

ตารางที่ 1 ฟังก์ชั่นจัดการไฟล์ Config

ฟังก์ชั่น	การทำงาน
config_read	ฟังก์ชั่นสำหรับโหลดข้อมูลไฟล์ proxy-fileter.conf ซึ่งเป็นไฟล์บอกรายละเอียด config ของระบบเข้าหน่วยความจำ 1. ACCESS_LOG 2. WHITE_LIST 3. BLACK_LIST1 และ BLACK_LIST2 4. DICTIONARY_1, DICTIONARY_2, DICTIONARY_3 และ DICTIONARY_4 5. PROXYPORT 6. LINE_CONSTANT_1, LINE_CONSTANT_2
config_print	ฟังก์ชั่นสำหรับพิมพ์รายละเอียดข้อมูลในไฟล์ proxy-filter.conf ออกทางหน้าจอ

ตารางที่ 2 ฟังก์ชั่นที่เกี่ยวข้องกับการจัดการ File Dicts

ฟังก์ชั่น	การทำงาน
dict_new	เซตค่า pointer dict ให้เป็นค่าว่าง และค่า count จำนวนคำใน dict ให้เป็น 0
dict_add	เก็บข้อมูล dict ในหน่วยความจำ โดยเก็บข้อมูลเป็นโหนดในลักษณะ linklist ที่มี pointer เป็นตัวชี้
dict_first	ใช้ pointer เพื่ออ่านข้อมูล dict ตั้งแต่โหนดแรก
dict_next	ใช้ pointer เพื่ออ่านโหนดข้อมูลถัดไป
dict_load	นำข้อมูล dict ที่เก็บในไฟล์เข้าหน่วยความจำผ่านฟังก์ชั่น dict_add
ฟังก์ชั่น	การทำงาน
dict_append, dict_remove	ทำการย้ายคำที่คาดว่าไม่เหมาะสมจาก dict3 ไป dict1 และจาก dict4 ไป dict2
dict_reset_frequency	เซตค่าความถี่สะสมของคำในแต่ละคำใน dict3 และ dict4 ให้เป็น 0
dict_frequency	แสดงค่าความถี่สะสมของคำแต่ละคำใน dict3 และ dict4 ผ่านทางหน้าจอ

ตารางที่ 3 ฟังก์ชั่นที่เกี่ยวข้องกับการเก็บข้อมูลเว็บไซต์

ฟังก์ชั่น	การทำงาน
weblist_exists	ฟังก์ชั่นกลางตรวจสอบ URL ว่ามีอยู่ในไฟล์ที่ต้องการตรวจสอบหรือไม่
weblist_add	ฟังก์ชั่นกลางในการเพิ่ม URL เข้าไปในไฟล์ที่ต้องการ
weblist_reset_black	สำหรับลบข้อมูลในไฟล์ blacklist1 และ blacklist2
weblist_reset_white	สำหรับลบข้อมูลในไฟล์ whitelist
weblist_reset	ลบข้อมูลใน blackist1, blactlist2 และ whitelist ผ่านฟังก์ชั่น weblist_reset_black, weblist_reset_white

ตารางที่ 4 ฟังก์ชั่นสำหรับทำกระบวนการ PCA

ฟังก์ชั่น	การทำงาน
eigen_output	เป็นฟังก์ชั่นหลักในการคำนวณหาค่า eigen
eigen_matrix	นำค่ามิติข้อมูลที่รวบรวมได้ไปแปลงให้อยู่ในรูป Matrix (1*8)
eigen_mius	คำนวณหาค่า DataAdjust โดยทำการนำมิติที่รวบรวมได้ลบด้วยค่าเฉลี่ยของแต่ละมิติ
eigen_multiply	คำนวณหาค่า FinalData โดยทำการนำค่า DataAdjust คูณกับค่า Eigenvector

ตารางที่ 5 ฟังก์ชั่นที่ใช้จัดการโปรแกรม Squid

ฟังก์ชั่น	การทำงาน
squid_read_log	อ่านข้อมูลการร้องขอของ Client จากไฟล์ access_log_path
squid_parser	ตรวจสอบข้อมูลจากที่ Client ร้องขอผ่าน squid_read_log ดังนี้ 1. ตรวจสอบว่าต้องเป็นการร้องขอจาก Client 2. ตรวจสอบว่าการร้องขอต้องมีชนิดของ Message เป็น 200 เป็นการร้องขอข้อมูลที่สำเร็จ
extract_http_address	ตัดข้อมูลที่ผ่านการตรวจสอบแล้วจากฟังก์ชั่น squid_parser ให้ได้ URL ที่ Client ร้องขอ
squid_reconfig	ส่งคำสั่ง “squid –k reconfigure” ให้ระบบทำการ reconfig ระบบ squid

ตารางที่ 6 ฟังก์ชั่นที่เกี่ยวข้องกับการจัดการเว็บไซต์

ฟังก์ชั่น	การทำงาน
httpclient	ติดต่อ proxy-server ส่ง URL ไปขอข้อมูล แบ่งข้อมูลออกเป็น 3 ส่วนคือ 1. result.tmp เป็นข้อมูลที่แสดงผลการติดต่อการร้องขอ 2. meta.tmp เป็นรายละเอียดข้อมูลเว็บไซต์ที่อยู่ใน TAG TITLE และ TAG META 3. httpclient.tmp เป็นรายละเอียดข้อมูลเว็บไซต์ที่อยู่ใน BODY ทั้งหมด
getnewlocation	หา url ใหม่ใน file result.tmp
httprequest	ตรวจสอบผลการร้องขอข้อมูลที่ได้จากฟังก์ชั่น httpclient กรณีเป็น โค๊ด 300 ให้ร้องขอใหม่ผ่าน getnewlocation
count_bad_word	อ่านข้อมูล httpclient เพื่อตรวจสอบว่ามีคำที่พบในแต่ละ dict ทั้งหมดกี่คำ
web_init	โหลด url ที่รับมาจากคำสั่งหรือคำร้องขอจาก client เข้า memory
web_read	ขอรายละเอียดข้อมูลเพื่อเก็บค่ามิติต่าง ๆทั้ง 8 มิติ

ตารางที่ 7 ฟังก์ชั่นที่เกี่ยวข้องกับการทำงานของ Proxy-Server

ฟังก์ชั่น	การทำงาน
init_dictionary	ฟังก์ชั่นหลักที่ใช้จัดการ Dict ผ่านฟังก์ชั่น dict_new, dict_load
record_web_list	ฟังก์ชั่นหลักที่ใช้จัดการเรื่องการเพิ่ม url ลงไฟล์ผ่านฟังก์ชั่น weblist_add
web_exists	ฟังก์ชั่นหลักในการตรวจสอบว่า url ที่เข้ามาในระบบมีอยู่ในไฟล์ใดบ้างโดยเริ่มตรวจสอบจากไฟล์ whitelist, blacklist1 และ blacklist2 ตามลำดับ ผ่านฟังก์ชั่น weblist_exists

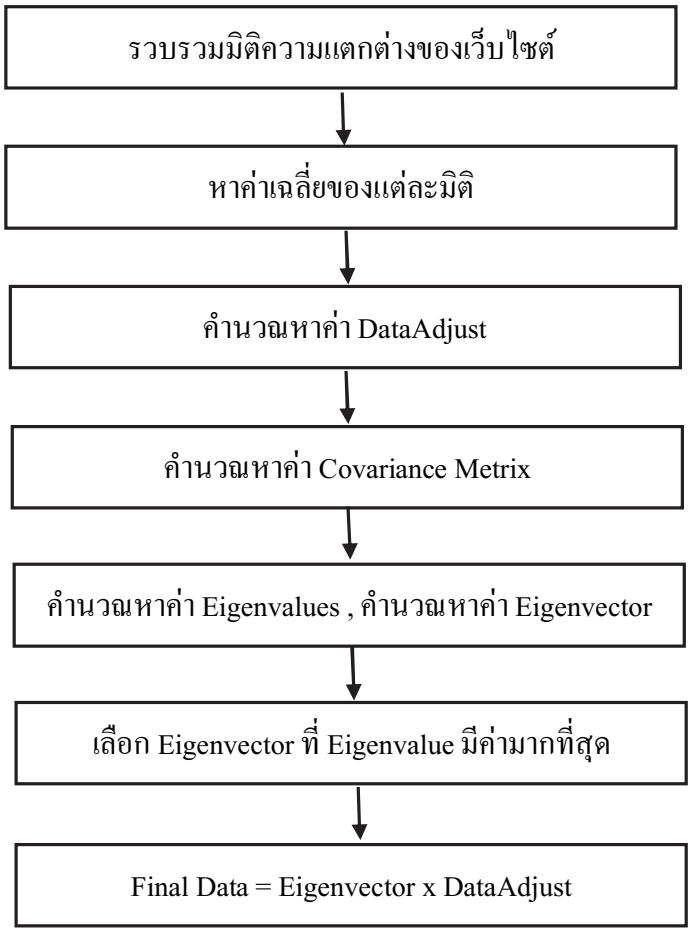
ตารางที่ 7 (ต่อ)

ฟังก์ชั่น	การทำงาน
move_word	ตรวจสอบคำที่คาดว่าไม่เหมาะสมใน dict3 คำใดต้องย้ายมาที่ dict1 และจาก dict4 มาที่ dict2 ผ่านฟังก์ชั่น dict_append, dic_add, dict_remove
check_web	1. ฟังก์ชั่น web_read ตรวจสอบว่า url นี้เหมาะสมหรือไม่จากทฤษฎี PCA 2. เขียน url ที่ไม่เหมาะสมลงไฟล์ blacklist1, blacklist2 3. ฟังก์ชั่น move_word ตรวจสอบคำที่คาดว่าไม่เหมาะสม
Daemon_filter	กรณีไม่มี argument เป็นการเรียก squid ทำงานเพื่อดึงข้อมูล url ที่ client ร้องขอเข้ามาตรวจสอบแบบอัตโนมัติ
action_by_param	กรณีมี argument ตรวจสอบว่าเป็นแบบใด 1. –s ฟังก์ชั่น web_init และฟังก์ชั่น check_web 2. –r ฟังก์ชั่น weblist_reset ให้ทำการลบข้อมูลใน blackist1, blactlist2 และ whitelist 3. –w ฟังก์ชั่น weblist_reset_white ให้ทำการลบข้อมูลทั้งหมดในไฟล์ whitelist
main	ฟังก์ชั่นหลักของโปรแกรม 1. ฟังก์ชั่น config_read จัดการ ไฟล์ config ของระบบ 2. ฟังก์ชั่น config_print แสดงรายละเอียด config ผ่านหน้าจอ 3. ฟังก์ชั่น init_dictionary โหลด dict ทั้ง 4 dict จากไฟล์เข้า memory 4. ตรวจสอบ argument ของคำสั่งร้องขอ

สร้างระบบการกรองเว็บไซต์ไม่เหมาะสม

นำกลุ่มคำไม่เหมาะสมและเกณฑ์การจำแนกเว็บไซต์ไม่เหมาะสมที่ได้มาสร้างระบบการกรองเว็บไซต์ไม่เหมาะสม โดยใช้ ทฤษฎี PCA (PCA : Principal Component Analysis) และนำไปใช้กับระบบงานจริง

หลักการวิเคราะห์แยกแยะส่วนประกอบ PCA (PCA :Principal Component Analysis)



ภาพที่ 12 ขั้นตอนการวิเคราะห์องค์ประกอบหลัก PCA (PCA :Principal Component Analysis)

หลักการวิเคราะห์องค์ประกอบหลัก

มีวิธีการทำดังนี้

1. รวบรวมมิติความแตกต่างของเว็บไซต์

โดยทำการเลือกปัจจัยที่เกี่ยวข้องกับเว็บไซต์ ซึ่งทำการรวบรวมมิติมาทั้งหมด 8 มิติที่บอกความแตกต่างของเว็บไซต์เพื่อใช้ในการทดสอบ

2. หาค่าเฉลี่ยของแต่ละมิติ (Mean)

ทำการหาค่าเฉลี่ยของข้อมูลแต่ละมิติที่ได้รวบรวมไว้

3. คำนวณหาค่า DataAdjust

คำนวณค่า Data Adjust โดยนำข้อมูลแต่ละมิติหักลบออกจากค่าเฉลี่ย (Mean) แต่ละมิติ

4. การคำนวณค่าเมทริกซ์โควาเรียนซ์ (Covariance Matrix)

นำค่า DataAdjust ที่ได้จากขั้นตอนที่ 3 มาหาค่าเมทริกซ์โควาเรียนซ์ โดยใช้สูตรโควาเรียนซ์ที่ได้กล่าวไปแล้ว จะได้เมทริกซ์โควาเรียนซ์ขนาด 8 x 8

5. การคำนวณไอแกนเวกเตอร์ (Eigenvectors) และไอแกนแวลูส์ (Eigenvalues) ของค่าเมทริกซ์โควาเรียนซ์

เมื่อค่าเมทริกซ์โควาเรียนซ์เป็น สี่เหลี่ยมจัตุรัสแล้ว(จำนวน Row เท่ากับจำนวนของ Column ดังนั้นจะสามารถคำนวณค่าของไอแกนเวกเตอร์ และไอแกนแวลูส์ สำหรับเมทริกซ์ได้

6. เลือกไอแกนเวกเตอร์ (Eigenvectors) ที่ ไอแกนแวลูส์ (Eigenvalues) มีค่ามากที่สุด

ขั้นตอนในการเลือก ไอแกนเวกเตอร์ (Eigenvectors) ที่ ไอแกนแวลูส์ (Eigenvalues) มีค่ามากที่สุด คือ การลดมิติหรือการเลือกไอแกนเวกเตอร์และไอแกนแวลูส์ที่มีค่ามากที่สุดที่เพียงพอต่อการอธิบายความแปรปรวนของกลุ่มได้ โดยมีกฎการเลือกดังนี้

6.1 ค่า ไอแกนแวลูส์ (Eigenvalues) > 1 โดยค่า Eigenvalue เป็นค่าที่บ่งบอกถึงความสามารถขององค์ประกอบว่าจะอธิบายความแปรปรวนของกลุ่มตัวแปรได้มากน้อยเพียงไร

6.2 ค่าความแปรปรวนสะสมไม่ควรต่ำกว่าร้อยละ 70 พิจารณาจากค่าความแปรปรวนโดยรวมของมิติที่เลือกมาว่ามีค่าความแปรปรวนรวมเพียงพอในการคัดกรองเว็บไซต์ไม่เหมาะสมหรือไม่ โดยค่าความแปรปรวนโดยรวมยิ่งมากหมายถึงสามารถอธิบายความแปรปรวนของกลุ่มได้มาก โดยการนำค่าไอแกนเวกเตอร์ที่เลือกมาจาก list ของไอแกนเวกเตอร์ และเปลี่ยนเป็นเมตริกซ์ให้ไอแกนเวกเตอร์อยู่ในคอลัมน์

$$\text{Eigenvector} = (\text{eig1 eig2 eig3} \ldots \text{eign}) \qquad (3.1)$$

7. คำนวณค่า Final Data

จะเป็นขั้นตอนสุดท้ายของ PCA และเป็นขั้นตอนที่ง่ายที่สุด โดยเราจะเลือกส่วนประกอบหรือ ปัจจัยที่ดีที่สุด (eigenvector) นำมา DotMatrix กับ DataAdjust เพื่อให้ได้ FinalData ที่เป็นข้อมูลตามมิติที่เรากำหนด

FinalData = Eigenvector x DataAdjust (3.2)

การประเมินผล

ประเมินผลความถูกต้องของระบบป้องกันการเข้าถึงเว็บไซต์ไม่เหมาะสม โดยคำนวณค่าร้อยละความถูกต้องในการกรองเว็บไซต์ไม่เหมาะสม ซึ่งการคำนวณค่าดังกล่าวเป็นดังแสดงในสมการ ดังนี้

$$\%\text{ความถูกต้อง} = \frac{\text{จำนวนเว็บไซต์ที่ให้ผลลัพธ์ได้ถูกต้อง}}{\text{จำนวนเว็บไซต์ที่ใช้ในการทดสอบ}}$$

บทที่ 4

ผลการดำเนินการวิจัย

ผลการดำเนินการวิจัย เรื่อง การพัฒนาระบบคัดกรองเว็บไซต์ที่ไม่เหมาะสมในระดับมัธยมศึกษา สามารถแบ่งได้เป็น 7 ขั้นตอนหลัก คือ

1. ศึกษาวิธีเขียนโปรแกรมภาษาซีบน Ubuntu9.04 ในการติดต่อขอข้อมูล html code จากเว็บไซต์

2. วิเคราะห์หากลุ่มคำเพื่อใช้จำแนกเว็บไซต์ไม่เหมาะสมออกจากเว็บไซต์ปกติ

3. วิเคราะห์หาจุดที่เหมาะสมที่ใช้ในการจำแนกเว็บไซต์ไม่เหมาะสมออกจากเว็บไซต์ปกติ โดยใช้ทฤษฎี PCA (PCA : Principal Component Analysis) และ ทฤษฎีการคำนวณทางสถิติ

4. สร้างระบบป้องกันการเข้าถึงเว็บไซต์ไม่เหมาะสม

5. วิเคราะห์การจำแนกเว็บไซต์โดยใช้ทฤษฎี SVM (SVM : Support Vector Machine) เพื่อเปรียบเทียบประสิทธิภาพการคัดกรอง

6. วิเคราะห์การเพิ่มกลุ่มคำไม่เหมาะสมแบบอัตโนมัติ

7. การทดสอบประสิทธิภาพการทำงาน

1. เขียนโปรแกรมภาษาซีบน Ubuntu9.04 ในการติดต่อขอข้อมูล html code

จากการศึกษาวิธีการเขียนโปรแกรมภาษาซี บน Ububtu9.04 ซึ่งใช้ gcc เป็นคอมไพล์เลอร์ (Compiler) เพื่อทำการแปลภาษา โดยใช้ vi เป็นเครื่องมือในการสร้างและแก้ไขซอร์สโค้ด (Source code) ในส่วนที่ทำการติดต่อกับ web server ประกอบไปด้วยชุดคำสั่ง ดังต่อไปนี้

```
host=gethostbyname(proxyserver)                เปลี่ยน hostname เป็นเลข IP
sockfd=socket(AF_INET,SOCK_STREM,0)    เปิด socket สำหรับเชื่อมต่อ
server.sin_family=AF_INET;                     สร้างการเชื่อมต่อ
server.sin_port=htons(config ->proxy_port);
memcpy(&server.sin_addr,host->h_addr_list[0], host->h_length);
memset(&(server.sin_zero),'\0',8);
connect(sockfd,(struct sockaddr*)&server,sizeof(struct sockaddr));
```

```
write(sockfd,”GET”,4);                                    ส่งคำร้องขอข้อมูล
write(sockfd,url,strlen(url));
write(sockfd,”HTTP/1.0\n\n”,11);
while ((numbeyte=read(sockfd,msgtemp,1)) ==1)   รับข้อมูลส่วนที่เป็น  header
{  if (msgtemp[0]!=’<’) { fputc(msgtemp[0],resultTemp);msgsize++;}
else { fputc(msgtemp[0],htmlTemp); msgsize=1; break;}}
while ((numbyte=read(sockfd,msgtemp,1)) ==1)    รับข้อมูลส่วนที่เป็น  html code
{ fputc(msgtemp[0],HtmlTemp); msgsize++;}
close(sockfd);                                            สิ้นสุดการเชื่อมต่อ
```

ข้อมูลที่ได้รับกลับมาจาก web server แบ่งเป็นส่วนของ header จะถูกเก็บลงไฟล์ result.tmp และในส่วนของ html code จะถูกเก็บลงไฟล์ httpclient.tmp

ข้อมูลที่ได้ใน result.tmp เป็นข้อมูลเกี่ยวกับผลการร้องขอข้อมูล โดยเฉพาะ html return code ซึ่งจะระบุว่าการร้องขอนั้นได้ผลลัพธ์อย่างไร ตัวอย่างข้อมูล มีดังนี้

ตัวอย่างที่ 1

HTTP/1.0 200 OK

Date: Wed,10 Dec 2009 08:07:05 GMT

Server: Apache

Content-Type: text/html

X-Cache: MISS from ns2

Proxy-Connection: close

แสดงว่าผลการร้องขอข้อมูลครั้งนี้ประสบผลสำเร็จ ได้รับข้อมูลตามที่ร้องขอ

ตัวอย่างที่ 2

HTTP/1.0 302 Moved Temporarily

Server: Resin/3.0.19

Location: http://www.spicevod.com/dispatcher/frontDoor

Content-Lenght: 82

Date: Mon,8 Dec 2009 04:27:54 GMT

X-Cache: MISS from ns2

X-Cache: MISS from com5

Proxy-Connection: close

แสดงว่าผลการร้องขอข้อมูลครั้งนี้ยังไม่เสร็จสมบูรณ์ ต้องทำการร้องขอไปยังปลายทางอื่นตามที่ระบุไว้ใน Location :

ข้อมูลที่ได้ใน httpclient.tmp เป็นรายละเอียดของ html code สำหรับการสร้างหน้า homepage ที่ browser จะนำไปใช้ ตัวอย่างข้อมูลมีดังนี้

<html>

<head>

<title>idol4u.com: รวมสุดยอด Album ภาพสวยๆ ที่หาได้ที่นี่ นักเรียน นักศึกษา netidol วัยรุ่น น่ารักใสๆ เพียบ!! </title>

<meta http-equiv=”Content=Type” content=”text/html;charset=”>

<meta http-equiv=”refresh” content=”0;URL=http://www.idol4u.com/index2.php”>

<META name=”description” content=”idol4u free gallery wallpaper pop idol and japanidol more!! And pic student narak narak free update more: รวมภาพนักเรียน รวมภาพนักศึกษา ดารานางแบบ บริการ sms”>

<META name=”keywords” content=”skin” solki,solki,solki,ringtone, ,download,logo,color,game,

java,chat,friend,msn,mobile,wallpaper,gallery, free,รวมภาพน่ารัก,รวมภาพ,สวยๆ,สาวๆ,น่ารักๆสาวๆ,sexy,idol,japan,ดารา,เกาหลี,นักเรียน,นักศึกษา,วัยรุ่น,board,fanclub,ฟรี,photo,photo2mobile,นางแบบ,สาวสวย,รูป,หาเพื่อน,หาแฟน,บอร์ด,เด็กๆน่ารัก,เด็กๆ,น้อง,โฟร์,เบเบ้,กุ๊กไก่,แพร์,นางแบบ,บอลลูน,เซียร์,เกมส์,ดูดวง,พอลล่า,solki, shinsolki,korea,นักร้อง,นักแสดง,กาหลี,วุ้นเส้น,น่ารักมาก,สาว,สวย,ใต้ดิน,โพสรูป,บอร์ด,เด็กสาว,เนียน,แอบถ่าย,คิขุ,ดารา,Britneyspear,msn,หางาน,รถยนต์,โพสภาพ,ฟรี,webboard,เด็กแนว,เด็กเส้น,บ้าน>

<link rel=”stylesheet” href=”textidol2.css” type=”text/css”>

<style type = “text/css”>

<

body { margin: 0px 0px; padding: 0px 0px}

a:link {color: #333333; text-decoration:none}

a:visited {color: #333333; text-decoration:none}

a:active {color: #ff6600; text-decoration:none}

```
a:hover {color: #ff6600; text-decoration:underline}
></style>
</head>
<body  leftmargin= “0”  topmargin = “0”  marginwidth = “0”  bgcolor= “#000000”>
<table  width= “100%”  border= “0” cellspacing= “0”  cellpadding = “0”  class =
“13pxmenu2” align= “center”>
<tr>
<td height= “614” valign= “top”>
<div  align= “center”>
<table  width= “100%”  border= “0” cellspacing= “0”  cellpadding = “0”  class =
“text13px5”  height= “192”>
</tr>
<td  height= “50”> </td>
</tr>
<tr>
<td  height= “197”>
<div  align= “center”><a href= “index2.php”> <font size = “4”><b><font  size = “3”
color= “#FF0000”>++
ไอดอลฟอร์ยูดอทคอม++ </font></b></font></a></div>
</td>
</tr>
<tr>
<td  height= “21”>
<div align= “center”>
</div>
</td>
</tr>
<td>
<div  align= “center”><span>
<script  language= “javascript1.1”>page= “home”;</script>
```

<script language= “javascript1.1”

Src= “http://truehits1.gits.net.th/data/i0016638.js”><img alt= ‘server uptime monitoring’ width=80 height=15 border=0 src=“http://ext.host-tracker.com/uptime-image/?s=15&t=112865 &m=”00.09&p=Total”><noscript>web server downtime monitoring </script>

</td>

</tr>

</table>

</td>

</tr>

</table>

</body>

</html>

การจำแนกเว็บไซต์ไม่เหมาะสมออกจากเว็บไซต์ปกตินั้น มีขั้นตอนเริ่มต้นจากการวิเคราะห์หากลุ่มคำไม่เหมาะสมและใช้หลักทฤษฎี PCA (PCA :Principal Component Analysis) เป็นเกณฑ์ในการจำแนกเว็บไซต์ที่ไม่เหมาะสมออกจากเว็บไซต์ปกติ ซึ่งแสดงดังแผนภาพ

ภาพที่ 13 ขั้นตอนการวิเคราะห์หากลุ่มคำไม่เหมาะสม ค่าทางสถิติ และทฤษฎี PCA (PCA : Principal Component Analysis)

2. วิเคราะห์หากลุ่มคำไม่เหมาะสมเพื่อใช้จำแนกเว็บไซต์ไม่เหมาะสมออกจากเว็บไซต์ปกติ

การเลือกกลุ่มคำที่ไม่เหมาะสมนั้นจะทำการแบ่งกลุ่มคำที่ไม่เหมาะสมเป็น 2 ประเภท ดังนี้

2.1 กลุ่มคำรุนแรง ยาเสพติด

วิธีเลือกกลุ่มคำรุนแรง ยาเสพติดที่ไม่เหมาะสม

2.1.1 รวบรวมกลุ่มคำรุนแรง ยาเสพติดที่ได้อ้างอิงจากวิทยานิพนธ์ “ความรุนแรงที่ปรากฏในการใช้ภาษาของข่าวอาชญากรรมในหนังสือพิมพ์รายวัน” ได้ดังนี้

มือปืน โหด วิปริต เชือดคอ ไอ้หื่น คลั่ง บุกยิง หื่นกาม ข่มขืน กาม เหยื่อ กระหน่ำแทง ฆาตกร ฟาด แทงพรุน จ้วงแทง กะซวก กะซวกแทง รุมฟัน เชือด ไส้ทะลัก ปาดคอ มีดกรีด ทุบด้วยของแข็ง กะโหลกร้าว ขยี้กาม ฆ่าตัวตาย กระสุน บุกยิง ยาอี ลั่นกระสุน แทงซ้ำ แทง จ่อยิง ข่มขืนยับ ไส้ไหล กรีดหน้า ปาดคอหอย ชำแหละ จ่อยิงหัว จ่อยิงขมับ รุมโทรม เสพยา บุกปล้ำ ไอ้โรคจิต ลวงข่มขืน รัดคอ เหี้ยมโหด ดิ่งตึก ใจโหด ปาด มั่วสุม ตีกัน เลือดเย็น กราดยิง กระหน่ำยิง มีดฟัน ขู่ฆ่า

2.1.2 นำกลุ่มคำที่รวบรวมได้ ทดสอบเพื่อหากลุ่มคำรุนแรง ยาเสพติด ที่มีอำนาจการแจกแจงสูงสุด เพื่อนำไปใช้ในการ Block เว็บไซต์ โดยวิธีการทดสอบนั้นมีขั้นตอนดังนี้

2.1.2.1หาจำนวนกลุ่มคำรุนแรง ยาเสพติดทั้งหมดในเว็บไซต์ตามกลุ่มคำรุนแรง ยาเสพติดที่ได้รวบรวมไว้ โดยทำการหากลุ่มคำรุนแรง ยาเสพติดจากเว็บไซต์ความรุนแรง ยาเสพติด 200 เว็บไซต์

2.1.2.2 หาจำนวนกลุ่มคำรุนแรง ยาเสพติดทั้งหมดในเว็บไซต์ตามกลุ่มคำรุนแรง ยาเสพติดที่ได้รวบรวมไว้ โดยทำการหากลุ่มคำรุนแรง ยาเสพติดจากเว็บไซต์ปกติ 200 เว็บไซต์

2.1.2.3 หาอัตราส่วนจำนวนกลุ่มคำรุนแรง ยาเสพติดที่พบในเว็บไซต์ความรุนแรง ยาเสพติด ต่อ จำนวนกลุ่มคำรุนแรง ยาเสพติดที่พบในเว็บไซต์ปกติ

ตารางที่ 8 อัตราส่วนจำนวนกลุ่มคำรุนแรง ยาเสพติดที่พบในเว็บไซต์ความรุนแรง ยาเสพติด ต่อ จำนวนกลุ่มคำรุนแรง ยาเสพติดที่พบในเว็บไซต์ปกติ

กลุ่มคำรุนแรง ยาเสพติด	**จำนวนเว็บไซต์ที่พบกลุ่มคำ**		**อัตราส่วนจำนวนกลุ่มคำรุนแรง ยาเสพติดที่พบในเว็บไซต์ความรุนแรงและยาเสพติดต่อจำนวนกลุ่มคำรุนแรง ยาเสพติดที่พบในเว็บไซต์ปกติ**
	จำนวนเว็บไซต์ความรุนแรง ยาเสพติดที่พบกลุ่มคำ	**จำนวนเว็บไซต์ปกติที่พบกลุ่มคำ**	
มือปืน	12	5	2.4000
โหด	76	5	15.2000

ตารางที่ 8 (ต่อ)

กลุ่มคำรุนแรง ยาเสพติด	จำนวนเว็บไซต์ที่พบกลุ่มคำ		อัตราส่วนจำนวนกลุ่มคำรุนแรงยาเสพติดที่พบในเว็บไซต์ความรุนแรงยาเสพติดต่อ จำนวนกลุ่มคำรุนแรงยาเสพติดที่พบในเว็บไซต์ปกติ
	จำนวนเว็บไซต์ความรุนแรงยาเสพติดที่พบกลุ่มคำ	จำนวนเว็บไซต์ปกติที่พบกลุ่มคำ	
วิปริต	25	4	6.2500
เชือดคอ	55	5	11.0000
ไอ้หื่น	15	3	5.0000
คลั่ง	14	4	3.5000
บุกยิง	41	4	10.2500
หื่นกาม	18	3	6.0000
ข่มขืน	72	4	18.0000
กาม	20	4	5.0000
เหยื่อ	45	7	6.4286
กระหน่ำแทง	35	7	5.0000
ฆาตกร	59	5	11.8000
ฟาด	70	6	11.6667
แทงพรุน	21	5	4.2000
จ้วงแทง	15	4	3.7500
กะซวก	30	4	7.5000
กะซวกแทง	20	3	6.6667
รุมฟัน	29	4	7.2500
เชือด	19	4	4.7500
ไส้ทะลัก	8	2	4.0000

ตารางที่ 8 (ต่อ)

กลุ่มคำรุนแรง ยาเสพติด	จำนวนเว็บไซต์ที่พบกลุ่มคำ		อัตราส่วนจำนวนกลุ่มคำรุนแรงยาเสพติดที่พบในเว็บไซต์ความรุนแรงยาเสพติดต่อ จำนวนกลุ่มคำรุนแรงยาเสพติดที่พบในเว็บไซต์ปกติ
	จำนวนเว็บไซต์ความรุนแรงยาเสพติดที่พบกลุ่มคำ	จำนวนเว็บไซต์ปกติที่พบกลุ่มคำ	
ปาดคอ	9	2	4.5000
มีดกรีด	13	7	1.8571
ทุบด้วยของแข็ง	11	5	2.2000
กะโหลกร้าว	21	4	5.2500
ขยี้กาม	6	2	3.0000
ฆ่าตัวตาย	72	5	14.4000
กระสุน	76	4	19.0000
บุกยิง	16	7	2.2857
ยาอี	54	8	6.7500
ลั่นกระสุน	45	6	7.5000
แทงซ้ำ	31	4	7.7500
แทง	98	4	24.5000
จ่อยิง	67	4	16.7500
ข่มขืนยับ	15	4	3.7500
ไส้ไหล	25	5	5.0000
กรีดหน้า	24	4	6.0000
ปาดคอหอย	8	2	4.0000
ชำแหละ	32	6	5.3333
จ่อยิงหัว	8	2	4.0000

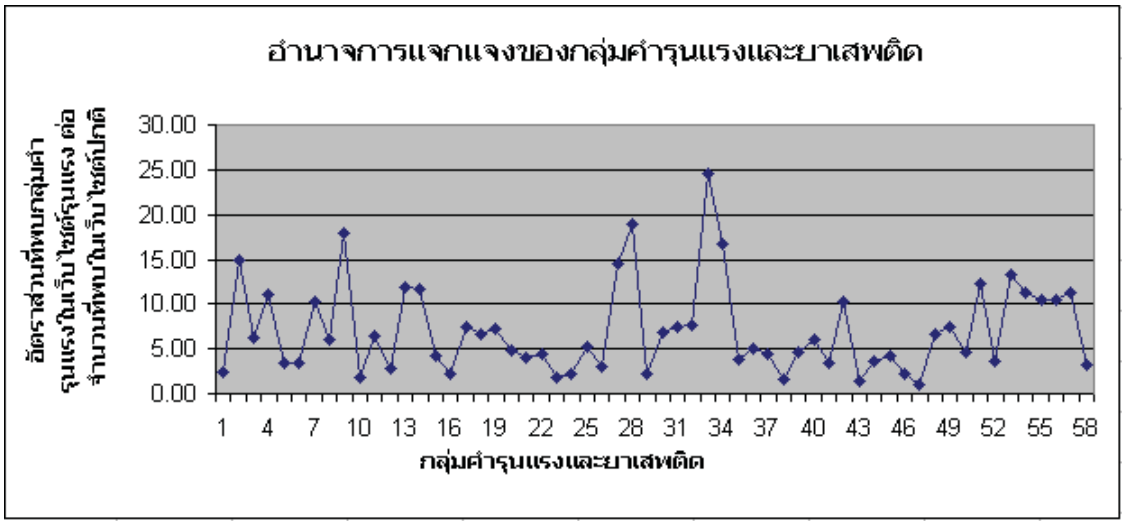
ตารางที่ 8 (ต่อ)

กลุ่มคำรุนแรง ยาเสพติด	จำนวนเว็บไซต์ที่พบกลุ่มคำ		อัตราส่วนจำนวนกลุ่มคำรุนแรงยาเสพติดที่พบในเว็บไซต์ความรุนแรงยาเสพติดต่อ จำนวนกลุ่มคำรุนแรงยาเสพติดที่พบในเว็บไซต์ปกติ
	จำนวนเว็บไซต์ความรุนแรงยาเสพติดที่พบกลุ่มคำ	จำนวนเว็บไซต์ปกติที่พบกลุ่มคำ	
จ่อยิงขมับ	10	3	3.3333
รุมโทรม	41	4	10.2500
เสพยา	28	11	2.5455
บุกปล้ำ	29	5	5.8000
ไอ้โรคจิต	17	4	4.2500
ลวงข่มขืน	15	7	2.1429
รัดคอ	15	6	2.5000
เหี้ยมโหด	12	2	6.0000
ดิ่งตึก	45	6	7.5000
ใจโหด	8	2	4.0000
ปาด	49	4	12.2500
มั่วสุม	29	8	3.6250
ตีกัน	66	5	13.2000
เลือดเย็น	34	3	11.3333
กราดยิง	21	2	10.5000
กระหน่ำยิง	42	4	10.5000
มีดฟัน	34	3	11.3333
ขู่ฆ่า	9	3	3.0000

2.1.2.4 จากตารางนำข้อมูลที่ได้ไปแสดงเป็นกราฟเพื่อหากลุ่มคำรุนแรง ยาเสพติดที่มี อำนาจการแจงแจงสูงสุด

แกน X แสดงถึง กลุ่มคำรุนแรง ยาเสพติด

แกน Y แสดงถึง อัตราส่วนจำนวนกลุ่มคำรุนแรง ยาเสพติดที่พบในเว็บไซต์ความรุนแรง ยาเสพติดต่อ จำนวนกลุ่มคำรุนแรง ยาเสพติดที่พบในเว็บไซต์ปกติ



ภาพที่ 14 กราฟแสดงอำนาจการแจงแจงของกลุ่มคำรุนแรง ยาเสพติด

2.1.2.5 จากกราฟทำให้ได้กลุ่มคำรุนแรง ยาเสพติดทั้งหมด 17 คำที่มีอำนาจการแจกแจงเริ่มที่ 10.25 ถึง 24.5 กลุ่มคำรุนแรง ยาเสพติดทั้งหมด 17 คำมีดังนี้

ตารางที่ 9 แสดงกลุ่มคำรุนแรง ยาเสพติดที่มีอำนาจการแจกแจงสูงสุด

กลุ่มคำรุนแรง ยาเสพติด	อำนาจการแจกแจง
แทง	24.50
กระสุน	19.00
ข่มขืน	18.00
จ่อยิง	16.75
โหด	15.20
ฆ่าตัวตาย	14.40

ตารางที่ 9 (ต่อ)

กลุ่มคำรุนแรง	อำนาจการแจกแจง
ตีกัน	13.20
ปาด	12.25
ฆาตกร	11.80
ฟาด	11.67
เลือดเย็น	11.33
มีดฟัน	11.33
เชือดคอ	11.00
กราดยิง	10.50
กระหน่ำยิง	10.50
บุกยิง	10.25
รุมโทรม	10.25

2.2 กลุ่มคำลามกอนาจาร

วิธีเลือกกลุ่มคำลามกอนาจารที่ไม่เหมาะสม

2.2.1 รวบรวมกลุ่มคำลามกอนาจารจากเว็บไซต์ที่มีเนื้อหา ข้อความลามกอนาจาร 150 เว็บไซต์ คำลามกอนาจารที่รวบรวมได้มีดังนี้

คำอนาจารที่1 คำอนาจารที่2 คำอนาจารที่3 คำอนาจารที่4 คำอนาจารที่5 คำอนาจารที่6 คำอนาจารที่7 คำอนาจารที่8 คำอนาจารที่9 คำอนาจารที่10 คำอนาจารที่11 คำอนาจารที่12 คำอนาจารที่13 คำอนาจารที่14 คำอนาจารที่15 คำอนาจารที่16 คำอนาจารที่17 คำอนาจารที่18 คำอนาจารที่19 คำอนาจารที่20 คำอนาจารที่21 คำอนาจารที่22 คำอนาจารที่23 คำอนาจารที่24 คำอนาจารที่25 คำอนาจารที่26 คำอนาจารที่27 คำอนาจารที่28 คำอนาจารที่29 คำอนาจารที่30 คำอนาจารที่31 คำอนาจารที่32 คำอนาจารที่33 คำอนาจารที่34 คำอนาจารที่35 คำอนาจารที่36 คำอนาจารที่37 คำอนาจารที่38 คำอนาจารที่39 คำอนาจารที่40 คำอนาจารที่41 คำอนาจารที่42 คำอนาจารที่43 คำอนาจารที่44 คำอนาจารที่45 คำอนาจารที่46 คำอนาจารที่47 คำอนาจารที่48 คำอนาจารที่49 คำอนาจารที่50 คำอนาจารที่51 คำอนาจารที่52 คำอนาจารที่53 คำอนาจารที่54 คำอนาจารที่55 คำอนาจารที่56 คำอนาจารที่57 คำอนาจารที่58 คำอนาจารที่59 คำอนาจารที่60 คำอนาจารที่61 คำอนาจารที่62

2.2.2. นำกลุ่มคำที่รวบรวมได้ ทดสอบเพื่อหากลุ่มคำลามกอนาจารที่มีอำนาจการแจกแจงสูงสุดเพื่อนำไปใช้ในการ Block เว็บไซต์โดยวิธีการทดสอบนั้นมีขั้นตอนดังนี้

2.2.2.1 หาจำนวนกลุ่มคำลามกอนาจารทั้งหมดในเว็บไซต์ ตามกลุ่มคำลามกอนาจารที่ได้รวบรวมไว้ โดยทำการหากลุ่มคำลามกอนาจารจากเว็บไซต์ลามกอนาจาร 150 เว็บไซต์

2.2.2.2 หาจำนวนกลุ่มคำลามกอนาจารทั้งหมดในเว็บไซต์ตามกลุ่มคำลามกอนาจารที่ได้รวบรวมไว้ โดยทำการหากลุ่มคำลามกอนาจารจากเว็บไซต์ปกติ 200 เว็บไซต์

2.2.2.3 หาอัตราส่วนจำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ลามกอนาจาร ต่อ จำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ปกติ

ตารางที่ 10 อัตราส่วนจำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ลามกอนาจาร ต่อ จำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ปกติ

กลุ่มคำลามกอนาจาร	**จำนวนเว็บไซต์ที่พบกลุ่มคำ**		**อัตราส่วนจำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ลามกอนาจาร ต่อ จำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ปกติ**
	จำนวนเว็บไซต์ลามกอนาจารที่พบกลุ่มคำ	**จำนวนเว็บไซต์ปกติที่พบกลุ่มคำ**	
คำอนาจารที่ 1	90	8	11.2500
คำอนาจารที่ 2	19	1	19.0000
คำอนาจารที่ 3	17	2	8.5000
คำอนาจารที่ 4	56	3	18.6667
คำอนาจารที่ 5	11	1	11.0000
คำอนาจารที่ 6	48	3	16.0000
คำอนาจารที่ 7	93	9	10.3333
คำอนาจารที่ 8	21	3	7.0000
คำอนาจารที่ 9	45	3	15.0000
คำอนาจารที่ 10	26	2	13.0000
คำอนาจารที่ 11	43	2	21.5000
คำอนาจารที่ 12	58	9	6.4444

ตารางที่ 10 (ต่อ)

กลุ่มคำลามกอนาจาร	จำนวนเว็บไซต์ที่พบกลุ่มคำ		อัตราส่วนจำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ลามกอนาจาร ต่อ จำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ปกติ
	จำนวนเว็บไซต์ลามกอนาจารที่พบกลุ่มคำ	จำนวนเว็บไซต์ปกติที่พบกลุ่มคำ	
คำอนาจารที่ 13	59	8	7.3750
คำอนาจารที่ 14	255	1	255.0000
คำอนาจารที่ 15	86	1	86.0000
คำอนาจารที่ 16	92	2	46.0000
คำอนาจารที่ 17	131	3	43.6667
คำอนาจารที่ 18	101	2	50.5000
คำอนาจารที่ 19	66	1	66.0000
คำอนาจารที่ 20	43	3	14.3333
คำอนาจารที่ 21	14	1	14.0000
คำอนาจารที่ 22	117	2	58.5000
คำอนาจารที่ 23	78	1	78.0000
คำอนาจารที่ 24	103	2	51.5000
คำอนาจารที่ 25	77	1	77.0000
คำอนาจารที่ 26	111	2	55.5000
คำอนาจารที่ 27	102	1	102.0000
คำอนาจารที่ 28	77	1	77.0000
คำอนาจารที่ 29	81	1	81.0000
คำอนาจารที่ 30	38	3	12.6667
คำอนาจารที่ 31	25	2	12.5000
คำอนาจารที่ 32	15	1	15.0000
คำอนาจารที่ 33	13	1	13.0000

ตารางที่ 10 (ต่อ)

กลุ่มคำลามกอนาจาร	จำนวนเว็บไซต์ที่พบกลุ่มคำ		อัตราส่วนจำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ลามกอนาจาร ต่อ จำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ปกติ
	จำนวนเว็บไซต์ลามกอนาจารที่พบกลุ่มคำ	จำนวนเว็บไซต์ปกติที่พบกลุ่มคำ	
คำอนาจารที่ 34	21	2	10.5000
คำอนาจารที่ 35	31	2	15.5000
คำอนาจารที่ 36	46	4	11.5000
คำอนาจารที่ 37	139	11	12.6364
คำอนาจารที่ 38	21	8	2.6250
คำอนาจารที่ 39	46	9	5.1111
คำอนาจารที่ 40	67	5	13.4000
คำอนาจารที่ 41	68	12	5.6667
คำอนาจารที่ 42	65	5	13.0000
คำอนาจารที่ 43	45	5	9.0000
คำอนาจารที่ 44	73	1	73.0000
คำอนาจารที่ 45	76	1	76.0000
คำอนาจารที่ 46	102	2	51.0000
คำอนาจารที่ 47	98	1	98.0000
คำอนาจารที่ 48	99	2	49.5000
คำอนาจารที่ 49	77	1	77.0000
คำอนาจารที่ 50	89	2	44.5000
คำอนาจารที่ 51	31	10	3.1000
คำอนาจารที่ 52	39	9	4.3333
คำอนาจารที่ 53	9	2	4.5000
คำอนาจารที่ 54	23	2	11.5000

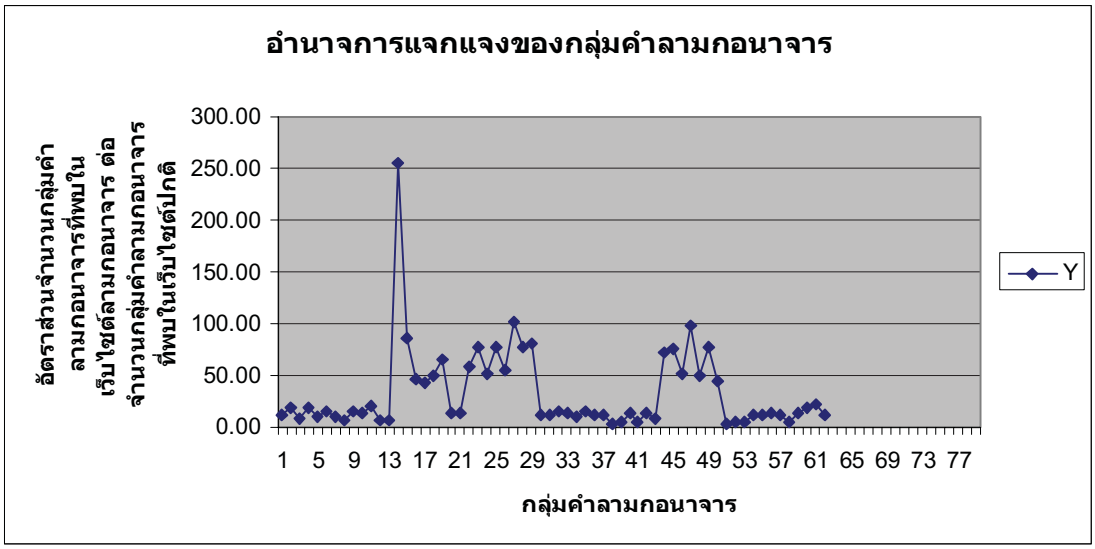
ตารางที่ 10 (ต่อ)

กลุ่มคำลามกอนาจาร	จำนวนเว็บไซต์ที่พบกลุ่มคำ		อัตราส่วนจำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ลามกอนาจาร ต่อ จำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ปกติ
	จำนวนเว็บไซต์ลามกอนาจารที่พบกลุ่มคำ	จำนวนเว็บไซต์ปกติที่พบกลุ่มคำ	
คำอนาจารที่ 55	36	3	12.0000
คำอนาจารที่ 56	43	3	14.3333
คำอนาจารที่ 57	74	6	12.3333
คำอนาจารที่ 58	54	9	6.0000
คำอนาจารที่ 59	26	2	13.0000
คำอนาจารที่ 60	19	1	19.0000
คำอนาจารที่ 61	22	1	22.0000
คำอนาจารที่ 62	59	5	11.8000

2.2.2.4 จากตารางนำข้อมูลที่ได้ไปแสดงเป็นกราฟเพื่อหากลุ่มคำลามกอนาจารที่มีอำนาจการแจงแจงสูงสุด

แกน X แสดงถึงกลุ่มคำลามกอนาจาร

แกน Y แสดงถึงอัตราส่วนจำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ลามกอนาจารต่อจำนวนกลุ่มคำลามกอนาจารที่พบในเว็บไซต์ปกติ



ภาพที่ 15 กราฟแสดงอำนาจการแจงแจงของกลุ่มคำลามกอนาจาร

2.2.2.5 จากกราฟทำให้ได้กลุ่มคำลามกอนาจารทั้งหมด 21 คำที่มีอำนาจการแจกแจงเริ่มที่ 43.67 ถึง 255 กลุ่มคำลามกอนาจารทั้งหมด 21 คำมีดังนี้

ตารางที่ 11 แสดงกลุ่มคำลามกอนาจารที่มีอำนาจการแจกแจงสูงสุด

กลุ่มคำลามกอนาจาร	อำนาจการแจกแจง
คำอนาจารที่ 14	255.00
คำอนาจารที่ 27	102.00
คำอนาจารที่ 47	98.00
คำอนาจารที่ 15	86.00
คำอนาจารที่ 29	81.00
คำอนาจารที่ 23	78.00
คำอนาจารที่ 25	77.00
คำอนาจารที่ 28	77.00
คำอนาจารที่ 49	77.00
คำอนาจารที่ 45	76.00
คำอนาจารที่ 44	73.00

ตารางที่ 11 (ต่อ)

กลุ่มคำลามกอนาจาร	อำนาจการแจกแจง
คำอนาจารที่ 19	66.00
คำอนาจารที่ 22	58.50
คำอนาจารที่ 26	55.50
คำอนาจารที่ 24	51.50
คำอนาจารที่ 46	51.00
คำอนาจารที่ 18	50.50
คำอนาจารที่ 48	49.50
คำอนาจารที่ 16	46.00
คำอนาจารที่ 50	44.50
คำอนาจารที่ 17	43.67

3. วิเคราะห์หาจุดที่เหมาะสมที่ใช้ในการจำแนกเว็บไซต์ไม่เหมาะสมออกจากเว็บไซต์ปกติโดยใช้ทฤษฎี PCA (PCA : Principal Component Analysis) และ ทฤษฎีการคำนวณทางสถิติ

3.1 กำหนดมิติที่ใช้ในการทดสอบจากเนื้อหาในเว็บไซต์

เพื่อใช้จำแนกความแตกต่างของข้อมูลในแต่ละเว็บไซต์ โดยทำการรวบรวมข้อมูลต่าง ๆ ได้ตามรูปแบบ 8 มิติ มิติความแตกต่างของเนื้อหาในเว็บไซต์ทั้ง 8 มิติมีดังนี้

3.1.1 จำนวนของ META TAG ซึ่งแสดงถึง Key word ที่สำคัญสำหรับการสืบค้นหาเว็บไซต์ของ Google และ description ของเว็บไซต์

3.1.2 จำนวนของ A HREF TAG ซึ่งแสดงถึงการ Link ไปยังเว็บไซต์อื่น

3.1.3 จำนวนของ IMG TAG ซึ่งแสดงถึงการแสดงรูปภาพ

3.1.4 จำนวนของ SCRIPT TAG ซึ่งเป็นโค๊ดสำหรับการตกแต่งเว็บไซต์

3.1.5 จำนวนคำที่ไม่เหมาะสมที่อยู่ใน META TAG และ TITLE TAG โดยทำการคำนวณคำที่ไม่เหมาะสมที่อยู่ในกลุ่มคำที่มีอำนาจการแจกแจงสูงสุด (จำนวน Pool)

3.1.6 จำนวนคำที่ไม่เหมาะสมที่อยู่ใน BODY โดยทำการคำนวณคำที่ไม่เหมาะสมที่อยู่ในกลุ่มคำที่มีอำนาจการแจกแจงสูงสุด

3.1.7 ค่าถ่วงน้ำหนักคำที่ไม่เหมาะสมใน META TAG และ TITLE TAG (ค่าถ่วงน้ำหนักของจำนวน Pool)

3.1.8 ค่าถ่วงน้ำหนักคำที่ไม่เหมาะสมใน BODY

3.2 ตรวจสอบความแตกต่างของเนื้อหาเว็บไซต์ทั้ง 8 มิติ กับกลุ่มตัวอย่างทดลอง

3.2.1 กลุ่มเว็บไซต์ลามกอนาจาร 150 เว็บไซต์

3.2.2 กลุ่มเว็บไซต์ความรุนแรง ยาเสพติด 200 เว็บไซต์

3.2.3 กลุ่มเว็บไซต์ปกติ 200 เว็บไซต์

ตารางที่ 12 แสดงคุณสมบัติของเนื้อหาเว็บไซต์

คุณสมบัติของเนื้อหาในเว็บไซต์	**เว็บไซต์ลามกอนาจาร**	**เว็บไซต์ความรุนแรงยาเสพติด**	**เว็บไซต์ปกติเทียบกับคำรุนแรงยาเสพติด**	**เว็บไซต์ปกติเทียบกับคำลามกอนาจาร**
จำนวนของ META TAG	773	1090	1106	1102
จำนวนของ A HREF TAG	13086	23623	31260	31288
จำนวนของ IMG TAG	4041	10234	21692	21715
จำนวนของ SCRIPT TAG	1445	3395	2329	2326
จำนวน Pool	998	458	0	0
จำนวนคำที่ไม่เหมาะสมที่อยู่ใน BODY	4360	2634	133	38
ค่าถ่วงน้ำหนักของจำนวน Pool	998	458	0	0
ค่าถ่วงน้ำหนักคำที่ไม่เหมาะสมใน BODY	4360	2634	133	38

3.3 วิเคราะห์จุดที่เหมาะสมในการแบ่งเว็บไซต์ลามกอนาจาร

กรณีเว็บไซต์ลามกอนาจาร จะทำการวิเคราะห์หาจุดที่เหมาะสมในการแบ่งเว็บไซต์ปกติออกจากเว็บไซต์ลามกอนาจาร โดยใช้ทฤษฎี PCA (PCA :Principal Component Analysis) และทฤษฎีการคำนวณทางสถิติ

ซึ่ง PCA เป็นทฤษฎีที่ทำการลดองค์ประกอบที่ไม่จำเป็นออกไป ดังนั้นจึงต้องทำการวิเคราะห์ก่อนว่าองค์ประกอบใดบ้างที่ไม่จำเป็นต่อระบบ โดยใช้ทฤษฎีความแปรปรวน หรือทฤษฎีการกระจายตัวของข้อมูลในแต่ละองค์ประกอบ เพื่อตัดสินใจว่าองค์ประกอบใดมีความสำคัญมากที่สุด และองค์ประกอบใดมีความสำคัญต่อระบบน้อยมาก เพื่อตัดองค์ประกอบที่ไม่สำคัญออกไปจากระบบ ขั้นตอนของ PCA มีดังนี้

3.3.1 ทำการหาค่าเฉลี่ยแต่ละองค์ประกอบหรือมิติในกลุ่มเว็บไซต์ลามกอนาจาร การหาค่าเฉลี่ยนั้นเพื่อนำไปใช้ในการคำนวณค่า Data Adjust ต่อไป

ตารางที่ 13 แสดงค่าเฉลี่ยคุณสมบัติของเว็บไซต์ลามกอนาจาร

คุณสมบัติของเว็บไซต์ลามกอนาจาร	**ค่าเฉลี่ยคุณสมบัติของเว็บไซต์ลามกอนาจาร**
จำนวนของ META TAG	5.1553
จำนวนของ A HREF TAG	87.24
จำนวนของ IMG TAG	26.94
จำนวนของ SCRIPT TAG	9.6933
จำนวนคำที่ไม่เหมาะสมที่อยู่ใน META TAG และ TITLE TAG	6.6533
จำนวนคำที่ไม่เหมาะสมที่อยู่ใน BODY	29.0667
ค่าถ่วงน้ำหนักของจำนวน Pool	6.6533
ค่าถ่วงน้ำหนักคำที่ไม่เหมาะสมใน BODY	29.0667

3.3.2 การปรับค่าข้อมูล (Data Adjust) คือ ขั้นตอนการนำข้อมูล Training เข้าระบบเพื่อให้ระบบได้ทำการวิเคราะห์ โดยข้อมูล Training ประกอบด้วยมิติหรือองค์ประกอบที่ต้องการทดสอบโดยนำข้อมูลในแต่ละองค์ประกอบลบกับค่าเฉลี่ยหรือค่ากลางของแต่ละองค์ประกอบ สาเหตุที่ต้องนำมิติมาปรับค่าโดยลบกับค่าเฉลี่ยก่อน เพราะระบบต้องทำการวิเคราะห์การกระจาย

ตัวขององค์ประกอบแต่ละองค์ประกอบดังนั้นการวัดการกระจายตัวจะใช้ทฤษฎีวาเรียนซ์ซึ่งเป็นทฤษฎีที่เราจะทำการกำหนดให้ข้อมูลที่เข้ามามีค่าเฉลี่ยรวมเป็น 0 เพื่อให้อยู่ในรูปการกระจายตัวของข้อมูลปกติ เพื่อไม่ให้ค่าเฉลี่ยมีผลต่อการวัดการกระจายตัวของข้อมูล ดังนั้นวิธีการที่ทำให้ค่าเฉลี่ยของข้อมูลเป็น 0 คือ การนำข้อมูลลบกับค่าเฉลี่ยก่อน ซึ่งเราเรียกขั้นตอนนี้ว่าการปรับข้อมูลหรือขั้นตอนการหาค่า Data Adjust โดยทำการปรับข้อมูลเว็บไซต์ลามกอนาจารทั้ง 150 เว็บไซต์

	1	2	3	4	5	6	7	8
1	-0.1533	7.76	-9.94	2.3667	-5.6533	-19.0667	-5.6533	-19.0667
2	-1.1533	-3.24	-23.94	-0.6333	-4.6533	-2.0667	-4.6533	-2.0667
3	-0.1533	-51.24	9.06	4.3667	-5.6533	-9.0667	-5.6533	-9.0667
4	-4.1533	-86.24	-20.94	-9.6333	-5.6533	0.9333	-5.6533	0.9333
5	-4.1533	-29.24	-25.94	-9.6333	-6.6533	-19.0667	-6.6533	-19.0667
6	1.8467	57.76	104.06	13.3667	-5.6533	-21.0667	-5.6533	-21.0667
7	-2.1533	-77.24	-21.94	13.3667	-6.6533	-16.0667	-6.6533	-16.0667
8	-0.1533	-51.24	7.06	-0.6333	-6.6533	-9.0667	-6.6533	-9.0667
9	-2.1533	-70.24	-21.94	-5.6333	-4.6533	5.9333	-4.6533	5.9333
10	-3.1533	-45.24	-7.94	-5.6333	-6.6533	-21.0667	-6.6533	-21.0667
11	-3.1533	35.76	21.06	-4.6333	-6.6533	-29.0667	-6.6533	-29.0667
12	1.8467	-74.24	-21.94	-5.6333	-6.6533	-18.0667	-6.6533	-18.0667
13	-3.1533	-53.24	-25.94	-0.6333	-6.6533	46.9333	-6.6533	46.9333
14	-1.1533	-14.24	-26.94	-2.6333	-5.6533	-11.0667	-5.6533	-11.0667
15	1.8467	67.76	118.06	13.3667	-6.6533	-16.0667	-6.6533	-16.0667

ภาพที่ 16 ตัวอย่างแสดงข้อมูลเว็บไซต์ลามกอนาจารที่มีการปรับค่าข้อมูล

3.3.3 ทำการคำนวณหาโควาเรียนซ์เมตริกซ์ของข้อมูลที่ปรับค่าแล้ว ขั้นตอนนี้จะเป็นขั้นตอนการวัดการกระจายตัวหรือความแปรปรวนของข้อมูล กรณีถ้าข้อมูลมีมิติเดียววิธีการวัดการกระจายตัวของข้อมูลคือการหาค่าวาเรียนซ์แต่ถ้ากรณีข้อมูลมีมากกว่า 1 มิติ ในทางสถิตินั้นการวัดการกระจายตัวของข้อมูลมากกว่า 1 มิติ จะทำการวัดการกระจายตัวของมิติตัวเองและต้องทำการวัดการกระจายตัวระหว่างมิติตัวเองกับมิติอื่น ๆ ด้วย เรียกว่าการหาค่า โควาเรียนซ์ โดยทำการวัดการกระจายตัวกับมิติอื่น ๆ เป็นคู่ ๆ และเนื่องจากมิติที่ทดสอบมีทั้งหมด 8 มิติ ดังนั้นเมื่อทำการหาค่า โควาเรียนซ์แล้ว ต้องทำการใส่ค่าความสัมพันธ์ โควาเรียนซ์ในแต่ละมิติลงเมตริกซ์ด้วย เพื่อใช้ในการคำนวณขั้นต่อไป โครงสร้างของโควาเรียนซ์เมตริกซ์ ทั้ง 8 มิติมีลักษณะความสัมพันธ์ดังนี้

$$\begin{pmatrix} cov(1,1) & cov(1,2) & cov(1,3) & cov(1,4) & cov(1,5) & cov(1,6) & cov(1,7) & cov(1,8) \\ cov(2,1) & cov(2,2) & cov(2,3) & cov(2,4) & cov(2,5) & cov(2,6) & cov(2,7) & cov(2,8) \\ cov(3,1) & cov(3,2) & cov(3,3) & cov(3,4) & cov(3,5) & cov(3,6) & cov(3,7) & cov(3,8) \\ cov(4,1) & cov(4,2) & cov(4,3) & cov(4,4) & cov(4,5) & cov(4,6) & cov(4,7) & cov(4,8) \\ cov(5,1) & cov(5,2) & cov(5,3) & cov(5,4) & cov(5,5) & cov(5,6) & cov(5,7) & cov(5,8) \\ cov(6,1) & cov(6,2) & cov(6,3) & cov(6,4) & cov(6,5) & cov(6,6) & cov(6,7) & cov(6,8) \\ cov(7,1) & cov(7,2) & cov(7,3) & cov(7,4) & cov(7,5) & cov(7,6) & cov(7,7) & cov(7,8) \\ cov(8,1) & cov(8,2) & cov(8,3) & cov(8,4) & cov(8,5) & cov(8,6) & cov(8,7) & cov(8,8) \end{pmatrix}$$

ภาพที่ 17 แสดงค่าความแปรปรวนระหว่างมิติตัวเองกับมิติอื่น ๆ

ซึ่งจะได้ โควาเรียนซ์เมตริกซ์ขนาด n x n (n คือจำนวนมิติทั้งหมด) จากข้อมูลการทดลองทำการคำนวณหาโควาเรียนซ์เมตริกซ์ของตัวอย่างข้อมูล8 มิติ ที่มีการปรับค่าแล้ว ซึ่งจะได้โควาเรียนซ์เมตริกซ์ขนาด 8x8 ดังนี้

	1	2	3	4	5	6	7	8
1	5743.4733	-3764.52	3781.38	345.4333	474.9733	-2192.5333	474.9733	-2192.5333
2	-3764.52	1291307.36	128721.16	7183.2	232022.48	165119.6	232022.48	165119.6
3	3781.38	128721.16	190236.46	14484.7	-8953.12	5820.6	-8953.12	5820.6
4	345.4333	7183.2	14484.7	8118.8333	-1956.0667	-3819.3333	-1956.0667	-3819.3333
5	474.9733	232022.48	-8953.12	-1956.0667	116089.9733	40476.4667	116089.9733	40476.4667
6	-2192.5333	165119.6	5820.6	-3819.3333	40476.4667	267401.3333	40476.4667	267401.3333
7	474.9733	232022.48	-8953.12	-1956.0667	116089.9733	40476.4667	116089.9733	40476.4667
8	-2192.5333	165119.6	5820.6	-3819.3333	40476.4667	267401.3333	40476.4667	267401.3333

ภาพที่ 18 แสดงผลการคำนวณหาโควาเรียนซ์ทั้ง 8 มิติที่ปรับค่าแล้วของเว็บไซต์ลามกอนาจาร

จากภาพแสดงให้เห็นถึงความสัมพันธ์การกระจายตัวของข้อมูลระหว่างมิติ เช่น เมื่อมิติที่ 1 กับมิติที่ 2 สามารถอธิบายความแปรปรวนหรือการกระจายตัวของข้อมูลได้ -3764.52 หรือ เมื่อมิติที่ 2 กับมิติที่ 5 สามารถอธิบายความแปรปรวนหรือการกระจายตัวของข้อมูลได้ 232022.48 โดยกระบวนการ PCA จะนำค่าความสัมพันธ์ของ ความแปรปรวนในแต่ละมิติ ไปวิเคราะห์เพื่อดึงค่าความแปรปรวนที่มากไปสร้างองค์ประกอบใหม่

3.3.4 คำนวณหาไอแกนเวกเตอร์และไอแกนแวลูส์ของโควาเรียนซ์เมตริกซ์

การวิเคราะห์องค์ประกอบหลัก เป็นเทคนิคการลดจำนวนตัวแปร

เทคนิคหนึ่ง โดยการสร้างตัวแปรใหม่เป็นฟังก์ชันเชิงเส้นของตัวแปรเดิม และตัวแปรใหม่จะต้องดึงรายละเอียดค่าความแปรปรวนจากตัวแปรเดิมมาไว้ตัวแปรใหม่ให้ได้มากที่สุด ซึ่งตัวแปรใหม่จะเขียนอยู่ในรูปองค์ประกอบหรือมิติเดิมคูณกับค่าไอเกนเวกเตอร์ดังนั้นค่าไอเกนเวกเตอร์คือค่าที่ทำให้เกิดการเปลี่ยนฟังก์ชันใหม่ ส่วนค่าไอเกนแวลูส์ คือ ค่าหรือจำนวนจริงที่บ่งบอกความสามารถของไอเกนเวกเตอร์ว่าจะอธิบายความแปรปรวนของกลุ่มตัวแปรได้มากน้อยเพียงใด

กำหนดให้ $(X_1, X_2, \ldots\ldots\ldots X_8)$ เป็นมิติทั้ง 8 ซึ่งมีเมตริกซ์ความแปรปรวนร่วม $\sum$ และมีไอเกนแวลูส์และไอเกนเวกเตอร์ $(\lambda_1, \lambda_1), (\lambda_2, \lambda_2), \ldots\ldots(\lambda_8, \lambda_8)$ เมื่อ λ คือไอเกนแวลูส์ โดยที่ $\lambda_1 \geq \lambda_2 \geq \ldots\ldots\ldots \geq \lambda_8$ ดังนั้นจะเขียนองค์ประกอบหลักหรือ องค์ประกอบใหม่ให้อยู่ในรูปฟังก์ชั่นเชิงเส้นของตัวแปร $X_1, X_2, \ldots\ldots\ldots X_8$ ได้ดังนี้

$$
\begin{aligned}
Y_1 &= \lambda_{11} X_1 + \lambda_{12} X_2 + \ldots. + \lambda_{1p} X_p \\
Y_2 &= \lambda_{21} X_1 + \lambda_{22} X_2 + \ldots. + \lambda_{2p} X_p \\
&\vdots \qquad\qquad\qquad \vdots \\
Y_p &= \lambda_{p1} X_1 + \lambda_{p2} X_2 + \ldots. + \lambda_{pp} X_p
\end{aligned}
\tag{4.1}
$$

ถ้าไอเกนเวกเตอร์หรือองค์ประกอบนั้นอธิบายความแปรปรวนของกลุ่มตัวอย่างได้น้อยกว่า 1 แล้ว ไอเกนแวลูส์นั้นก็ไม่มีประโยชน์ที่จะนำองค์ประกอบนั้นมาใช้ หากตัวแปรที่นำมาวิเคราะห์มีจำนวนน้อย การวิเคราะห์อาจให้ผลเป็นองค์ประกอบแค่ 2-3 องค์ประกอบเท่านั้น ถ้าหากองค์ประกอบที่นำมาวิเคราะห์มีจำนวนมากอาจจะได้องค์ประกอบจำนวนมาก แต่เราอาจกำหนดเกณฑ์อื่น ๆ ในการเลือกองค์ประกอบได้ เช่น ค่าไอเกนแวลูส์ต้องมากกว่า 1 ซึ่งเป็นเกณฑ์ที่กำหนดไว้ในทุก ๆ โปรแกรมคอมพิวเตอร์อยู่แล้ว

	1
1	1462219.2993
2	471275.6799
3	208845.3002
4	107585.1833
5	6901.1873
6	5562.0899
7	1.8159e-010
8	-6.9671e-012

ภาพที่ 19 แสดงค่า Eigenvalues ที่คำนวณได้

	1	2	3	4	5	6	7	8
1	-0.0026	-0.0051	0.0114	0.036	0.0174	-0.9991	7.5048e-016	-5.878e-018
2	0.9254	-0.2535	0.0518	-0.2767	-0.001	-0.0105	2.1242e-016	8.188e-017
3	0.0927	-0.0881	0.8248	0.5445	-0.0785	0.0279	-3.056e-016	1.4361e-017
4	0.004	-0.0178	0.0668	0.0379	0.9967	0.0196	1.1141e-016	-1.4995e-015
5	0.1858	-0.012	-0.3911	0.5586	0.0037	0.0153	-0.2178	-0.6727
6	0.1816	0.6809	0.057	-0.0058	0.0079	-0.0034	0.6727	-0.2178
7	0.1858	-0.012	-0.3911	0.5586	0.0037	0.0153	0.2178	0.6727
8	0.1816	0.6809	0.057	-0.0058	0.0079	-0.0034	-0.6727	0.2178

ภาพที่ 20 แสดงค่า Eigenvectors ที่คำนวณได้

จากสมการ $Y_p = \lambda_{p1} X_1 + \lambda_{p2} X_2 + \ldots. + \lambda_{pp} X_p$

ซึ่งเป็นองค์ประกอบใหม่ จะได้ว่า $\sum_{i=1}^{p} Var(x_i) = \lambda_1 + \lambda_2 + \ldots\ldots\ldots + \lambda_8 = \sum_{i=1}^{p} Var(y_i)$ จากสมการจะเห็นได้ว่าผลรวมความแปรปรวนของทุก ๆ ตัวแปร ในเมตริกซ์ข้อมูล X และผลรวมของความแปรปรวนของทุก ๆ องค์ประกอบใหม่จะให้ค่าเท่ากันและเท่ากับผลรวมของไอเกนแวลูส์ ซึ่งสัดส่วนความแปรปรวนของข้อมูลในแต่ละองค์ประกอบใหม่ จะสามารถแสดงได้ดังนี้ สัดส่วนความแปรปรวนรวมของประชากรขององค์ประกอบใหม่ที่ k คือ

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots\ldots + \lambda_p} \quad ; \quad k = 1, 2, \ldots\ldots p \tag{4.2}$$

ดังนั้นถ้าองค์ประกอบตัวที่ k มีสัดส่วนความแปรปรวนมาก แสดงว่าองค์ประกอบดังกล่าวสามารถอธิบายความผันแปรของตัวแปรเดิมได้มาก และเนื่องจาก $\lambda_1 \geq \lambda_2 \geq \ldots\ldots\ldots \geq \lambda_8 \geq 0$ แสดงว่าความแปรปรวนขององค์ประกอบใหม่ที่สอดคล้องกับ λ ตัวแรก ๆ จะมีค่าที่สูง ด้วยเหตุนี้จึงใช้หลักการดังกล่าวในการพิจารณาหาจำนวนขององค์ประกอบที่เหมาะสมสำหรับการลดมิติของข้อมูลต่อไป

3.3.5 การลดมิติโดยเลือกองค์ประกอบที่สำคัญ

ขั้นตอนในการลดมิตินั้นจะทำการพิจารณาไอเกนเวกเตอร์และไอเกนแวลูส์ โดยทำการเลือกค่าไอเกนเวกเตอร์ที่มีค่าไอแกนเวลูส์สูงที่สุด ซึ่งแสดงถึงความสัมพันธ์ของข้อมูลที่มีความสำคัญ ดังนั้นหากทำการเรียงข้อมูลตามค่าของไอเกนแวลูส์จากมากไปหาน้อยจะได้องค์ประกอบที่เรียงลำดับความสำคัญ ทำให้สามารถตัดสินใจที่จะละเลยองค์ประกอบที่มี

ความสำคัญน้อยกว่า ซึ่งจะทำให้สูญเสียข้อมูลบางส่วน โดยเป็นการลดมิติที่ทำให้ข้อมูลสูญเสียไม่มาก แต่มีผลทำให้สามารถนำข้อมูลที่เหลือเฉพาะส่วนที่สำคัญไปประมวลผลต่อได้ง่ายและเร็วขึ้น

โดยกฎที่ใช้ในการบ่งบอกว่าจะลดขนาดมิติเท่าไหร่ถึงจะเหมาะสมคือค่า Eigenvalue > 1 (Cooley and Lohnes 1971; Kim and Mueller 1978; Kerlinger 1986; Stevens 1996) โดยค่า Eigenvalue เป็นค่าที่บ่งบอกถึงความสามารถขององค์ประกอบว่าจะอธิบายความแปรปรวนของกลุ่มตัวแปรได้มากน้อยเพียงไร โดยปกติถ้าองค์ประกอบนั้นอธิบายความแปรปรวนของกลุ่มตัวอย่างได้น้อยกว่า 1 ค่า Eigenvalue นั้นก็ไม่มีประโยชน์ที่จะนำองค์ประกอบนั้นมาใช้พิจารณาในทฤษฎี PCA และกฎอีกข้อหนึ่งที่ใช้พิจารณาในการลดขนาดมิตินั้น คือ จะพิจารณาจากค่าความแปรปรวนโดยรวมของมิติที่เลือกมาว่ามีค่าความแปรปรวนรวมเพียงพอในการคัดกรองเว็บไซต์ไม่เหมาะสมหรือไม่ โดยค่าความแปรปรวนโดยรวมยิ่งมากหมายถึงสามารถอธิบายความแปรปรวนของกลุ่มได้มาก

โดยปกติแล้วค่าความแปรปรวนสะสมไม่ควรต่ำกว่าร้อยละ 70 (Cooley and Lohnes 1971; Kim and Mueller 1978; Kerlinger 1986; Stevens 1996)

จากผลการทดลองนำค่า Eigenvalues และความแปรปรวนในแต่ละมิติ ที่ได้มาพิจารณา ได้ผลดังตารางต่อไปนี้

ตารางที่ 14 แสดงค่า Eigenvalue ความแปรปรวน และความแปรปรวนรวม ที่คำนวณได้

Eigenvector	**Eigenvalues**	**%variance**	**Cumulative % of variance**
1	1462219.2993	64.63	64.63
2	471275.6799	20.83	85.46
3	208845.3002	9.23	94.69
4	107585.1833	4.76	99.45
5	6901.1873	0.31	99.75
6	5562.0899	0.25	100
7	1.82E-10	8.03E-15	100
8	-6.97E-12	-3.1E-16	100

จากทฤษฎีทางสถิติความแปรปรวนหมายถึง ถ้าองค์ประกอบที่เลือกมามีค่าความแปรปรวนหรือสามารถวัดค่าการกระจายตัวของข้อมูลน้อย แสดงว่าองค์ประกอบนั้นไม่สามารถอธิบายได้ว่าข้อมูลทั้ง 2 นั้นแตกต่างกันอย่างไร แต่ถ้าองค์ประกอบนั้นมีค่าความแปรปรวนมาก หรือสามารถวัดค่าการกระจายตัวของข้อมูลได้มาก แสดงว่าองค์ประกอบนั้นสามารถอธิบายความแตกต่างของข้อมูล 2 กลุ่ม และแบ่งแยกข้อมูล 2 กลุ่มนั้นออกจากกันได้

ทฤษฎีของ PCA ต้องทำการเรียงลำดับค่าไอเกนแวลูส์จากมากไปน้อยดังนั้นลำดับของไอเกนเวกเตอร์ตัวแรกต้องอธิบายความแปรปรวนได้มากที่สุด

จากตาราง สามารถสรุปกรณีการเลือกมิติที่ลดรูปแล้วในแต่ละค่าที่แตกต่างกันได้ดังนี้

กรณีลดลงเหลือ 2 มิติสามารถอธิบายความแปรปรวนรวมได้ 85.46 เปอร์เซ็นต์

กรณีลดลงเหลือ 3 มิติสามารถอธิบายความแปรปรวนรวมได้ 94.69 เปอร์เซ็นต์

กรณีลดลงเหลือ 4 มิติสามารถอธิบายความแปรปรวนรวมได้ 99.45 เปอร์เซ็นต์

กรณีลดลงเหลือ 5 มิติสามารถอธิบายความแปรปรวนรวมได้ 99.75 เปอร์เซ็นต์

กรณีลดลงเหลือ 6-8 มิติสามารถอธิบายความแปรปรวนรวมได้ 100.00 เปอร์เซ็นต์

การทดสอบจะทำการเลือก Eigenvalues 2 มิติแรก เพราะค่า Eigenvalues 2 มิติแรกมีค่า Eigenvalues > 1 และค่าความแปรปรวนรวนรวม 85.46 เปอร์เซ็นต์ ซึ่ง 2 มิติแรกสามารถอธิบายความแปรปรวนของกลุ่มตัวอย่างได้ 85.46 เปอร์เซ็นต์ ซึ่งมีความแปรปรวนรวมมากกว่า 70% แล้ว เพื่อนำไปใช้ในการวิเคราะห์ข้อมูลเชิงเส้น ซึ่งเป็นข้อมูลเพียงพอต่อการคัดกรองเว็บไซต์แล้ว

ดังนั้นจากข้อมูลการทดสอบมี 8 มิติ จะทำการลดมิติของข้อมูลให้เหลือ 2 มิติเพื่อนำไปใช้ในการวิเคราะห์ข้อมูลเชิงเส้น

	1	2
1	-0.0026	-0.0051
2	0.9254	-0.2535
3	0.0927	-0.0881
4	0.004	-0.0178
5	0.1858	-0.012
6	0.1816	0.6809
7	0.1858	-0.012
8	0.1816	0.6809

ภาพที่ 21 แสดงค่า Eigenvectors ที่มีค่า Eigenvalues สูงสุด 2 อันดับ

3.3.6 ทำการหาสมการ FinalData สำหรับการทดสอบข้อมูลเว็บไซต์

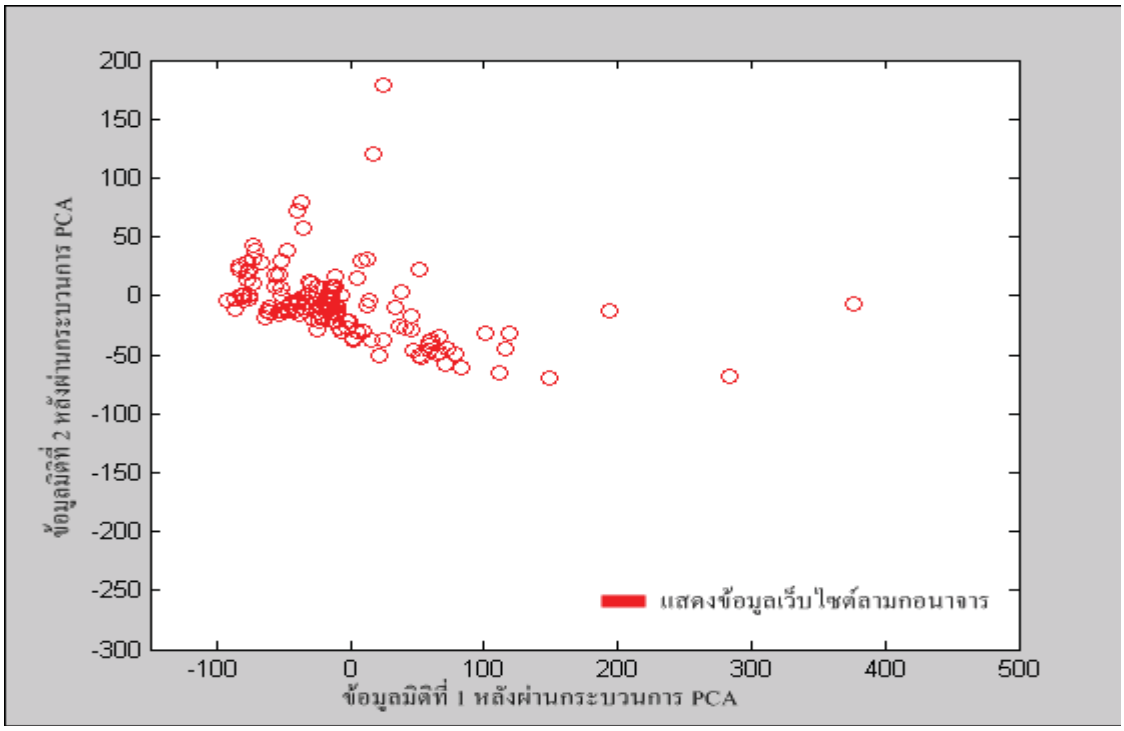
$$FinalData = eigenvectors * DataAdjust \quad (4.3)$$

เมื่อ eigenvectors มีขนาดเมทริกซ์ 8*2

DataAdjust ของ 1 เว็บไซต์ลามกอนาจาร มีขนาดเมทริกซ์ 1*8

ดังนั้น FinalData มีขนาดเมทริกซ์ 1*2 (ค่า X,Y)

3.3.7 นำข้อมูลเว็บไซต์ลามกอนาจาร 150 เว็บไซต์เข้าสมการเพื่อหาค่า FinalData ของแต่ละเว็บไซต์ ดังนั้นจะได้ข้อมูลเมทริกซ์ 1*2 หรือค่า (X,Y) จำนวน 150 ค่าเพื่อนำไป Plot กราฟเพื่อแสดงตำแหน่งของข้อมูลแต่ละเว็บไซต์



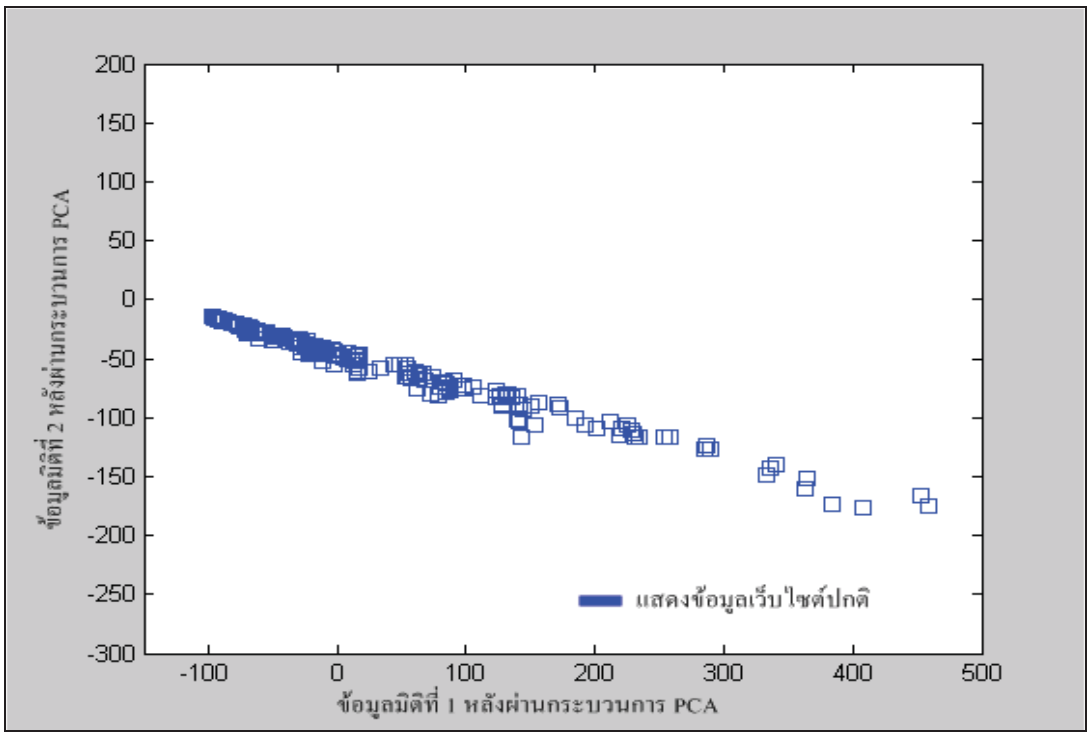
ภาพที่ 22 แสดงตำแหน่งของข้อมูลแต่ละเว็บไซต์ของเว็บไซต์ลามกอนาจาร

3.3.8 คำนวณค่ามิติทั้ง 8 มิติกับเว็บไซต์ปกติทั้ง 200 เว็บไซต์ (จำนวนคำไม่เหมาะสมคำนวณจากประเภทคำลามกอนาจาร) และนำไปคำนวณหาค่า DataAdjust ของข้อมูลโดยใช้ค่าเฉลี่ยของเว็บไซต์ลามกอนาจาร

	1	2	3	4	5	6	7	8
1	-1.1533	-47.24	73.06	15.3667	-6.6533	-29.0667	-6.6533	-29.0667
2	-0.1533	-62.24	-19.94	-4.6333	-6.6533	-29.0667	-6.6533	-29.0667
3	0.8467	-30.24	51.06	-7.6333	-6.6533	-29.0667	-6.6533	-29.0667
4	-2.1533	-87.24	-26.94	-9.6333	-6.6533	-29.0667	-6.6533	-29.0667
5	-3.1533	142.76	77.06	0.3667	-6.6533	-29.0667	-6.6533	-29.0667
6	-2.1533	-16.24	-15.94	-7.6333	-6.6533	-29.0667	-6.6533	-29.0667
7	-3.1533	110.76	87.06	4.3667	-6.6533	-29.0667	-6.6533	-29.0667
8	0.8467	179.76	35.06	-6.6333	-6.6533	-29.0667	-6.6533	-29.0667
9	2.8467	602.76	371.06	1.3667	-6.6533	-29.0667	-6.6533	-29.0667
10	-0.1533	251.76	156.06	6.3667	-6.6533	-29.0667	-6.6533	-29.0667
11	-1.1533	84.76	196.06	18.3667	-6.6533	-29.0667	-6.6533	-29.0667
12	2.8467	499.76	102.06	11.3667	-6.6533	-29.0667	-6.6533	-29.0667
13	-2.1533	132.76	339.06	-7.6333	-6.6533	-29.0667	-6.6533	-29.0667
14	5.8467	-70.24	17.06	10.3667	-6.6533	-29.0667	-6.6533	-29.0667
15	12.8467	183.76	123.06	8.3667	-6.6533	-23.0667	-6.6533	-23.0667

ภาพที่ 23 ตัวอย่างแสดงข้อมูลเว็บไซต์ปกติที่มีการปรับค่าข้อมูล

3.3.9 นำข้อมูล Data Adjust เข้าสูตร FinalData โดยใช้ค่า Eigenvectors ของเว็บไซต์ลามกอนาจารดังนั้นจะได้ข้อมูลเมทริกซ์ 1*2 หรือค่า (X,Y) จำนวน 200 ค่าเพื่อนำไป Plot กราฟเพื่อแสดงตำแหน่งของข้อมูลแต่ละเว็บไซต์



ภาพที่ 24 แสดงตำแหน่งของข้อมูลแต่ละเว็บไซต์ของเว็บไซต์ปกติ

3.3.10 นำข้อมูลของกราฟทั้ง 2 นำมารวมกันเพื่อดูตำแหน่งของข้อมูลแต่ละเว็บไซต์และทำการคำนวณหาจุดที่เหมาะสมในการแบ่งเว็บไซต์ปกติออกจากเว็บไซต์ลามกอนาจาร โดยใช้สมการเส้นตรง Y = MX + C เป็นตัวแบ่งข้อมูลออกจากกัน ทฤษฎีการหาสมการเส้นตรงนั้น จะหาจากจุด 2 จุดคือ จุดต่ำสุดของเว็บไซต์ลามกอนาจารและจุดสูงสุดของเว็บไซต์ปกติที่ใช้ทดสอบลามกอนาจาร

x1=-96.2639 y1= -14.5071 จุดสูงสุดของเว็บไซต์ปกติประเภทลามกอนาจาร

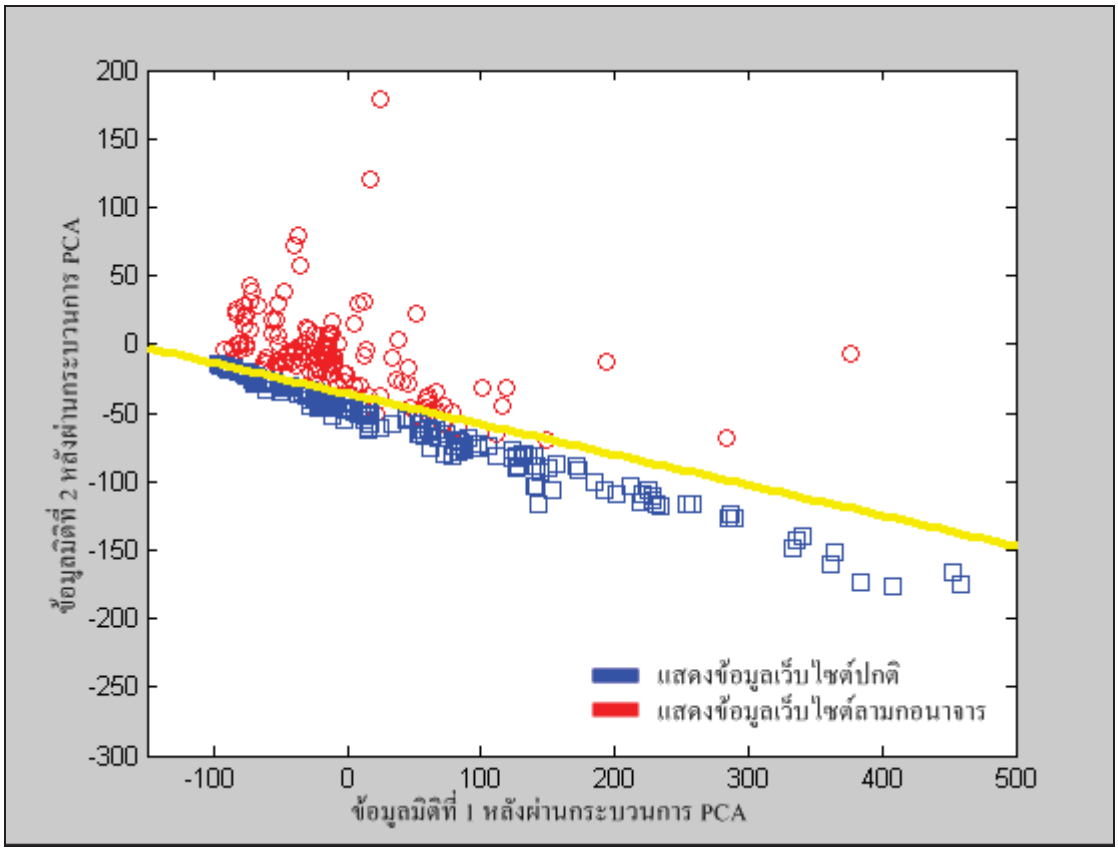
x2= 149.1764 y2= -69.0588 จุดต่ำสุดของเว็บไซต์ลามกอนาจาร

ค่าความชัน M ที่คำนวณได้คือ -0.2223

ค่าคงที่ C ที่คำนวณได้คือ -35.9028

กำหนดสมการให้อยู่ในรูป y = mx + b หรือ AX + BY + C = 0

สมการเส้นตรง Y = -0.2223X -35.9028 หรือ 0.2223X – Y + 35.9028 = 0



ภาพที่ 25 แสดงการแบ่งเว็บไซต์ลามกอนาจารออกจากเว็บไซต์ปกติ

3.3.11 การทดสอบเมื่อมีข้อมูลเว็บไซต์เข้ามาในระบบจะคำนวณ DataAdjust ทั้ง 8 มิติโดยใช้ค่าเฉลี่ยแต่ละมิติของเว็บไซต์ลามกอนาจาร และค่า Eigenvector ของเว็บไซต์ลามกอนาจารและทำการคำนวณค่า FinalData ให้ได้ค่า (x, y) ของเว็บไซต์ และทำการนำค่า (x,y) ที่ได้เข้าสูตรสมการเส้นตรงหาค่า C และนำค่า C ที่ได้ไปตรวจสอบว่าเว็บไซต์ที่เข้ามาเป็นเว็บไซต์ลามกอนาจารหรือไม่โดยพิจารณาจากเงื่อนไขต่อไปนี้

ค่า C > -35.9028 แสดงว่าข้อมูลอยู่เหนือสมการเส้นตรงซึ่งหมายความว่าเป็นเว็บไซต์ลามกอนาจาร

ค่า C ≤ -35.9028 แสดงว่าข้อมูลอยู่ใต้สมการเส้นตรงซึ่งหมายความว่าเป็นเว็บไซต์ปกติ

3.4 วิเคราะห์จุดที่เหมาะสมในการแบ่งเว็บไซต์ความรุนแรง ยาเสพติด

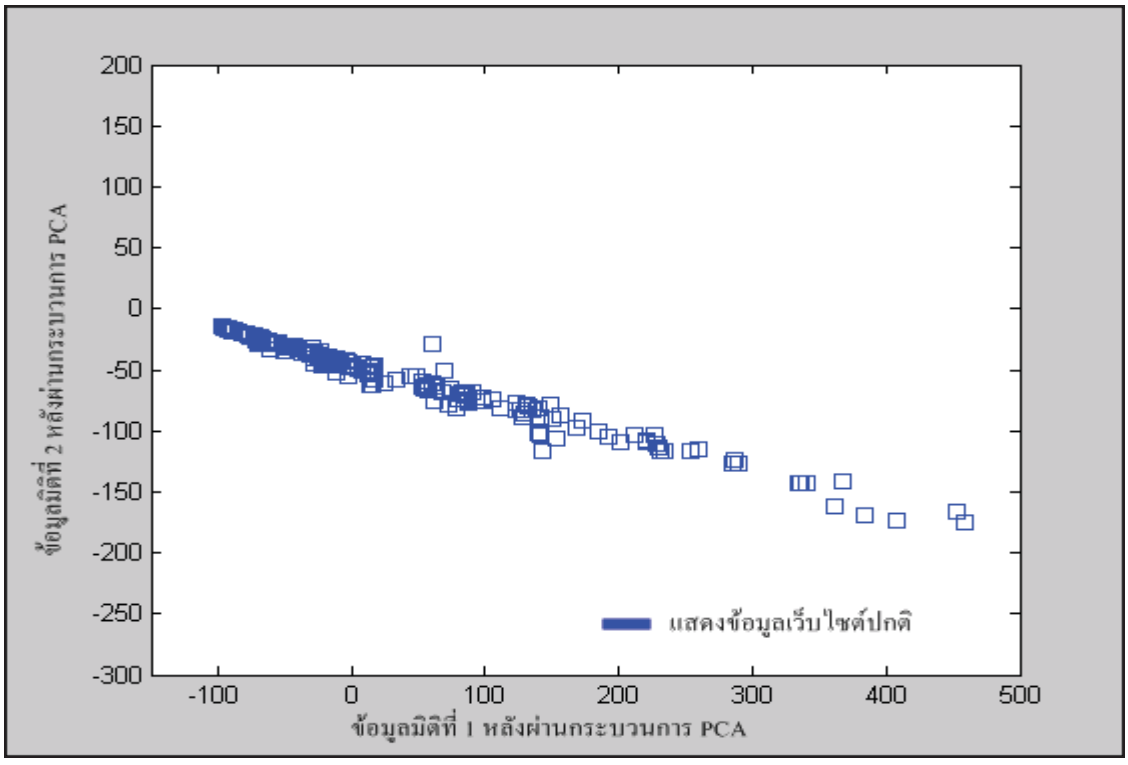
กรณีเว็บไซต์ความรุนแรง ยาเสพติด จะทำการวิเคราะห์หาจุดที่เหมาะสมในการแบ่งเว็บไซต์ปกติออกจากเว็บไซต์รุนแรง ยาเสพติดโดยใช้ทฤษฎี PCA (PCA : Principal Component Analysis) และ ทฤษฎีการคำนวณทางสถิติดังนี้ โดยการทดสอบการแบ่งเว็บไซต์ความรุนแรง ยาเสพติดออกจากเว็บไซต์ปกตินั้นจะใช้ Eigenvector และค่าเฉลี่ยของเว็บไซต์ลามกอนาจารเป็นตัวแบ่งข้อมูลเพราะค่า FinalData ที่คำนวณได้สามารถแบ่งข้อมูลเว็บไซต์ปกติและเว็บไซต์ความรุนแรง ยาเสพติดได้ในระดับที่น่าพอใจ

3.4.1 การปรับค่าข้อมูล (Data Adjust) ของเว็บไซต์ปกติที่จะทดสอบความรุนแรง ยาเสพติด คือ ทำการลบข้อมูลทุกๆค่าในมิตินั้นๆด้วยค่าเฉลี่ยของมิติ นั้นๆ (ค่าเฉลี่ยของเว็บไซต์ลามกอนาจาร) โดยทำการปรับข้อมูลเว็บไซต์ปกติทั้ง 200 เว็บไซต์

	1	2	3	4	5	6	7	8
1	-1.1533	-47.24	73.06	15.3667	-6.6533	-29.0667	-6.6533	-29.0667
2	-0.1533	-62.24	-19.94	-4.6333	-6.6533	-29.0667	-6.6533	-29.0667
3	0.8467	-30.24	51.06	-7.6333	-6.6533	-29.0667	-6.6533	-29.0667
4	-2.1533	-87.24	-26.94	-9.6333	-6.6533	-29.0667	-6.6533	-29.0667
5	-3.1533	142.76	77.06	0.3667	-6.6533	-29.0667	-6.6533	-29.0667
6	-2.1533	-16.24	-15.94	-7.6333	-6.6533	-27.0667	-6.6533	-27.0667
7	-3.1533	110.76	87.06	4.3667	-6.6533	-29.0667	-6.6533	-29.0667
8	0.8467	179.76	35.06	-6.6333	-6.6533	-29.0667	-6.6533	-29.0667
9	2.8467	602.76	371.06	1.3667	-6.6533	-29.0667	-6.6533	-29.0667
10	-0.1533	251.76	156.06	6.3667	-6.6533	-29.0667	-6.6533	-29.0667
11	-1.1533	84.76	196.06	18.3667	-6.6533	-23.0667	-6.6533	-23.0667
12	2.8467	499.76	102.06	11.3667	-6.6533	-29.0667	-6.6533	-29.0667
13	-2.1533	132.76	339.06	-7.6333	-6.6533	-29.0667	-6.6533	-29.0667
14	5.8467	-70.24	17.06	10.3667	-6.6533	-29.0667	-6.6533	-29.0667
15	12.8467	183.76	123.06	8.3667	-6.6533	-29.0667	-6.6533	-29.0667

ภาพที่ 26 ตัวอย่างแสดงข้อมูลเว็บไซต์ปกติที่ใช้ทดสอบความรุนแรง ยาเสพติดที่มีการปรับค่าข้อมูล

3.4.2 โดยนำข้อมูล Data Adjust เข้าสูตร FinalData โดยใช้ค่า eigenvectors ของเว็บไซต์ลามกอนาจารดังนั้นจะได้ข้อมูลเมตริกซ์ 1*2 หรือค่า (X,Y) จำนวน 200 ค่าเพื่อนำไป Plot กราฟเพื่อแสดงตำแหน่งของข้อมูลแต่ละเว็บไซต์



ภาพที่ 27 แสดงตำแหน่งของข้อมูลแต่ละเว็บไซต์ของเว็บไซต์ปกติ

3.4.3 นำข้อมูลของกราฟทั้ง 2 นำมารวมกันเพื่อดูตำแหน่งของข้อมูลแต่ละเว็บไซต์และทำการคำนวณหาจุดที่เหมาะสมในการแบ่งเว็บไซต์ปกติออกจากเว็บไซต์ความรุนแรงยาเสพติด

ทฤษฎีการหาสมการเส้นตรงนั้น จะหาจากจุด 2 จุดคือ จุดต่ำสุดของเว็บไซต์ลามกอนาจารและจุดสูงสุดของเว็บไซต์ปกติที่ใช้ทดสอบความรุนแรง

x1= -96.2690 y1= -14.5174 จุดสูงสุดของเว็บไซต์ปกติประเภทความรุนแรงยาเสพติด

x2= 149.1764 y2= -69.0588 จุดต่ำสุดของเว็บไซต์ลามกอนาจาร

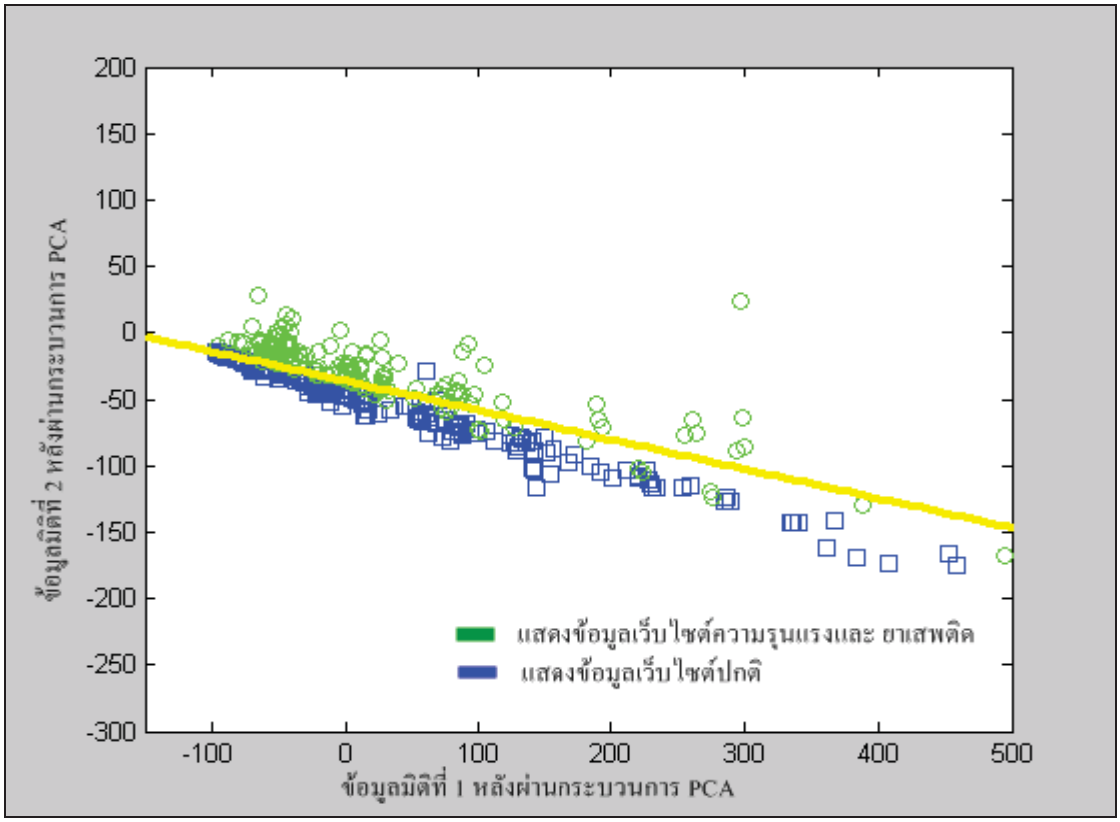
ค่าความชัน M ที่คำนวณได้คือ -0.2221

ค่าคงที่ C ที่คำนวณได้คือ -35.9201

กำหนดสมการให้อยู่ในรูป $y = mx + b$ หรือ $AX + BY + C = 0$

สมการเส้นตรง $Y = -0.2221X - 35.9201$ หรือ $0.2221X - Y + 35.9201 = 0$

และทำการทดสอบนำข้อมูลเว็บไซต์รุนแรงเข้ามาในระบบเพื่อดูตำแหน่งของข้อมูลแต่ละเว็บไซต์ความรุนแรง



ภาพที่ 28 แสดงการแบ่งเว็บไซต์ความรุนแรง ยาเสพติดออกจากเว็บไซต์ปกติ

3.4.4 การทดสอบเมื่อมีข้อมูลเว็บไซต์เข้ามาในระบบจะทำการคำนวณ DataAdjust ทั้ง 8 มิติโดยใช้ค่าเฉลี่ยแต่ละมิติของเว็บไซต์ลามกอนาจาร และค่า Eigenvector ของเว็บไซต์ลามกอนาจารและทำการคำนวณค่า FinalData ให้ได้ค่า (x, y) ของเว็บไซต์ และทำการนำค่า (x,y) ที่ได้เข้าสูตรสมการเส้นตรงหาค่า C และนำค่า C ที่ได้ไปตรวจสอบว่าเว็บไซต์ที่เข้ามาเป็นเว็บไซต์ความรุนแรงหรือไม่ โดยพิจารณาจากเงื่อนไขต่อไปนี้

ค่า C > -35.9201 แสดงว่าข้อมูลอยู่เหนือสมการเส้นตรงซึ่งหมายความว่าเป็นเว็บไซต์ความรุนแรง ยาเสพติด

ค่า C ≤ -35.9201 แสดงว่าข้อมูลอยู่ใต้สมการเส้นตรงซึ่งหมายความว่าเป็นเว็บไซต์ปกติ

3.5 ผลการทดสอบการแบ่งกลุ่มเว็บไซต์โดยใช้ทฤษฎี PCA (PCA : Principal Component Analysis)

เป็นขั้นตอนของการนำผลเว็บไซต์ที่ทำการแบ่งประเภทเว็บไซต์แล้วมาประเมินประสิทธิภาพ โดยข้อมูลทดสอบการคัดกรองเว็บไซต์เป็นข้อมูลชุดเดียวกับที่ทดสอบ SVM

1. ทดสอบเว็บไซต์ความรุนแรง ยาเสพติด แบ่งเป็นเว็บไซต์ปกติ 100 เว็บไซต์ และเว็บไซต์ความรุนแรง ยาเสพติด 100 เว็บไซต์

2. ทดสอบเว็บไซต์ลามกอนาจาร แบ่งเป็นเว็บไซต์ปกติ 100 เว็บไซต์ และเว็บไซต์ลามกอนาจาร 100 เว็บไซต์

ทำการตรวจสอบดูว่าเว็บไซต์ที่แบ่งประเภทนั้นมีค่าเป็นอย่างไร โดยทำการประเมินประสิทธิภาพค่าต่าง ๆ ตามตารางดังนี้

ตารางที่ 15 แสดงประสิทธิภาพการคัดกรองเว็บไซต์ด้วยวิธี PCA (PCA : Principal Component Analysis)

ประเภทเว็บไซต์	ค่าความถูกต้อง Accuracy	ค่าความแม่นยำ Precision	ค่าความระลึก Recall
เว็บไซต์ความรุนแรง ยาเสพติด	89.5	96.47	82
เว็บไซต์ลามกอนาจาร	94.5	100	89

3.6 การประเมินประสิทธิภาพจำนวนเว็บไซต์ที่ได้จากการคำนวณสมการเส้นตรง

การประเมินประสิทธิภาพของสมการเส้นตรงทั้ง 2 วิธีนั้น ทำการทดสอบว่า Model ที่สร้างนั้นมีความผิดพลาดในการแบ่งข้อมูลมากน้อยเพียงใดโดยคิดเป็นอัตราส่วนร้อยละของข้อมูลทั้งหมด

ตารางที่ 16 แสดงประสิทธิภาพสมการเส้นตรงสำหรับการแบ่งประเภทเว็บไซต์ความรุนแรงยาเสพติดกับเว็บไซต์ปกติ

ประเภทเว็บไซต์	จำนวนเว็บไซต์	จำแนกถูก		จำแนกผิด	
		จำนวนเว็บ	ร้อยละ	จำนวนเว็บ	ร้อยละ
เว็บไซต์ความรุนแรงยาเสพติด	200	174	87	26	13
เว็บไซต์ปกติประเภทความรุนแรง ยาเสพติด	200	197	98.5	3	1.5
เฉลี่ยรวม	400	371	92.75	29	7.25

ตารางที่ 17 แสดงประสิทธิภาพสมการเส้นตรงสำหรับการแบ่งประเภทเว็บไซต์ลามกอนาจารกับเว็บไซต์ปกติ

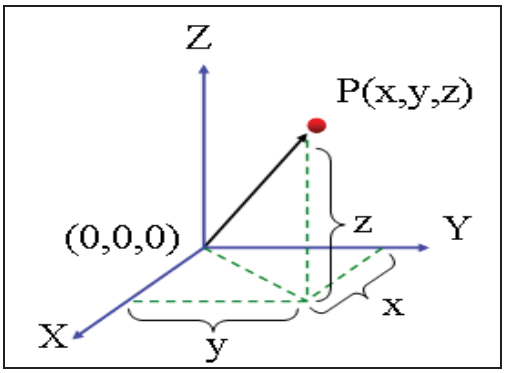
ประเภทเว็บไซต์	จำนวนเว็บไซต์	จำแนกถูก		จำแนกผิด	
		จำนวนเว็บ	ร้อยละ	จำนวนเว็บ	ร้อยละ
เว็บไซต์ลามกอนาจาร	150	143	95.33	7	4.67
เว็บไซต์ปกติลามกอนาจาร	200	200	100	0	0
เฉลี่ยรวม	350	343	98	7	2

สรุปประสิทธิภาพของสมการเส้นตรงที่ใช้แบ่งข้อมูล จากการทดลองพบว่ามีจำนวนเว็บไซต์ไม่เหมาะสม หลุดเข้ามาในกลุ่มเว็บไซต์ปกติ หลังจากพิจารณาดูแล้วพบว่าเว็บไซต์ที่หลุดเข้ามานั้นเป็นเว็บไซต์ที่มีคำที่ไม่เหมาะสมน้อยกว่าเกณฑ์ที่กำหนดไว้ สาเหตุมาจากเว็บไซต์

นั้นมีการหลบเลี่ยงไปใช้คำอื่น ที่สื่อความหมายใกล้เคียงกัน ระบบจึงไม่สามารถตรวจสอบได้ กรณีที่มีเว็บไซต์ปกติหลุดเข้ามาใน กลุ่มของเว็บไซต์ไม่เหมาะสมเป็นเพราะว่าเว็บไซต์นั้นมีคำไม่เหมาะสมอยู่เพียงพอที่ระบบตรวจสอบแล้วตัดสินใจให้เว็บไซต์นั้นเป็นเว็บไซต์ไม่เหมาะสม

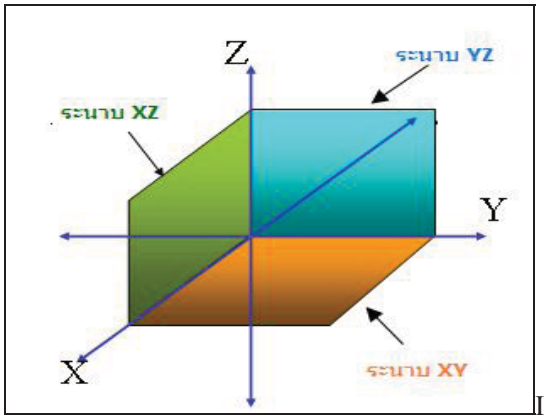
3.7 ทดสอบวิเคราะห์องค์ประกอบ 3 มิติ

การบอกตำแหน่งใน 3 มิติ เรานิยมใช้ Cartesian Coordinate ซึ่งประกอบด้วยเลข 3 ตัว (x,y,z) ซึ่งแทนระยะทางตามแนวแกน X, Y และ Z ตามลำดับ วัดเทียบกับจุด Origin (0,0,0)



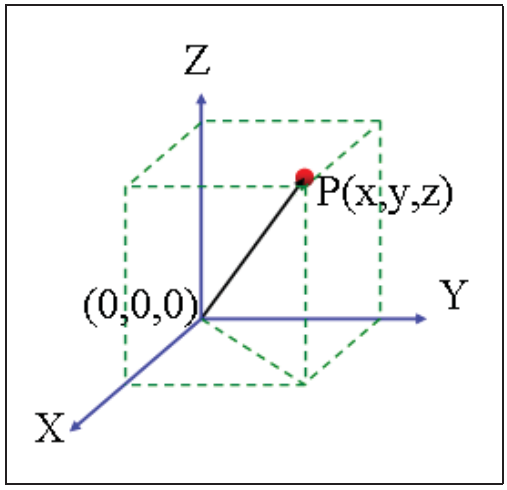
ภาพที่ 29 แสดงระยะทางตามแนวแกน X Y และ Z เทียบกับจุด Origin

จากแกนพิกัดทั้ง 3 แกน จะทำให้เกิดระนาบ 3 ระนาบ ได้แก่ ระนาบ xy ระนาบ xz และระนาบ yz



ภาพที่ 30 แสดงระนาบทั้ง 3 ระนาบบนระบบพิกัดฉาก 3 มิติ

Cartesian Coordinate มีชื่อเรียกอีกชื่อว่า Rectangular Coordinate เนื่องจากว่า แกน X, Y, Z ตั้งฉากกันดังนั้นจุด (x,y,z) ก็คือจุดยอดของกล่องสี่เหลี่ยม (Rectangular box) จุดที่อยู่ตรงข้ามกับจุด (0,0,0)



ภาพที่ 31 แสดงการบอกพิกัดในระนาบ 3 มิติ

หลังจากได้ข้อมูลบนพิกัดฉาก 3 มิติแล้วทำการพิจารณาเพื่อหาทางแบ่งข้อมูลอาจใช้สมการมิติใด ๆ เพื่อทำการแบ่งข้อมูล 2 กลุ่มออกจากกัน สมการเชิงเส้นสามารถมีตัวแปรได้มากกว่า 2 ตัว สมการเชิงเส้นทั่วไปที่มีจำนวนตัวแปร n ตัวสามารถเขียนได้ในรูปแบบ

$$a_1 x_1 + a_2 x_2 + \ldots + a_n x_n = \mathrm{b} \qquad (4.4)$$

ซึ่ง $a_1, a_2, \ldots, a_n$ เป็นสัมประสิทธิ์ $x_1, x_2, \ldots, x_n$ คือตัวแปร และ b คือค่าคงตัว เมื่อเราต้องการเขียนสมการตัวแปรน้อยๆ เช่น 3 ตัว เราอาจแทนที่ x_1, x_2, x_3 ด้วยชื่อตัวแปรอื่นๆ เช่น x, y, z ได้ตามต้องการ

3.8 การประยุกต์ใช้ทฤษฎีอื่นในการแบ่งข้อมูล

กรณีข้อมูลที่ได้ไม่สามารถแบ่งได้ด้วยเส้นตรงอาจใช้ทฤษฎี K-Means ซึ่งเป็นอัลกอริทึมที่ใช้แยกข้อมูลออกเป็นกลุ่มๆ โดยอาศัยคุณลักษณะต่างๆเพื่อแยกข้อมูลออกเป็น K กลุ่ม การจัดกลุ่มแบบ K-Means มีขั้นตอนดังนี้

3.8.1 แบ่งกลุ่มข้อมูลออกเป็น K กลุ่ม ซึ่งไม่ใช่เซตว่าง กรณีนี้กำหนด K = 2 เพื่อทำการแบ่งข้อมูลเว็บไซต์ออกเป็น 2 กลุ่ม

3.8.2 เริ่มแรกในการทดลองจะทำการ กำหนดจุดเซ็นทรอยด์ 2 จุด โดยทำการเลือกจุดข้อมูลใดๆในชุดข้อมูลนั้น

3.8.3 นำข้อมูลเทียบกับจุดกึ่งกลางทั้ง 2 จุด เพื่อกำหนดกลุ่มให้กับข้อมูล โดยเลือกระยะจากข้อมูลไปจุดกึ่งกลางแต่ละจุดที่ให้ระยะทางใกล้ที่สุด สูตรการหาระยะทางจะใช้การคำนวณระยะทางแบบยูคลิดีน (Euclidean distance) โดยมีสมการดังนี้

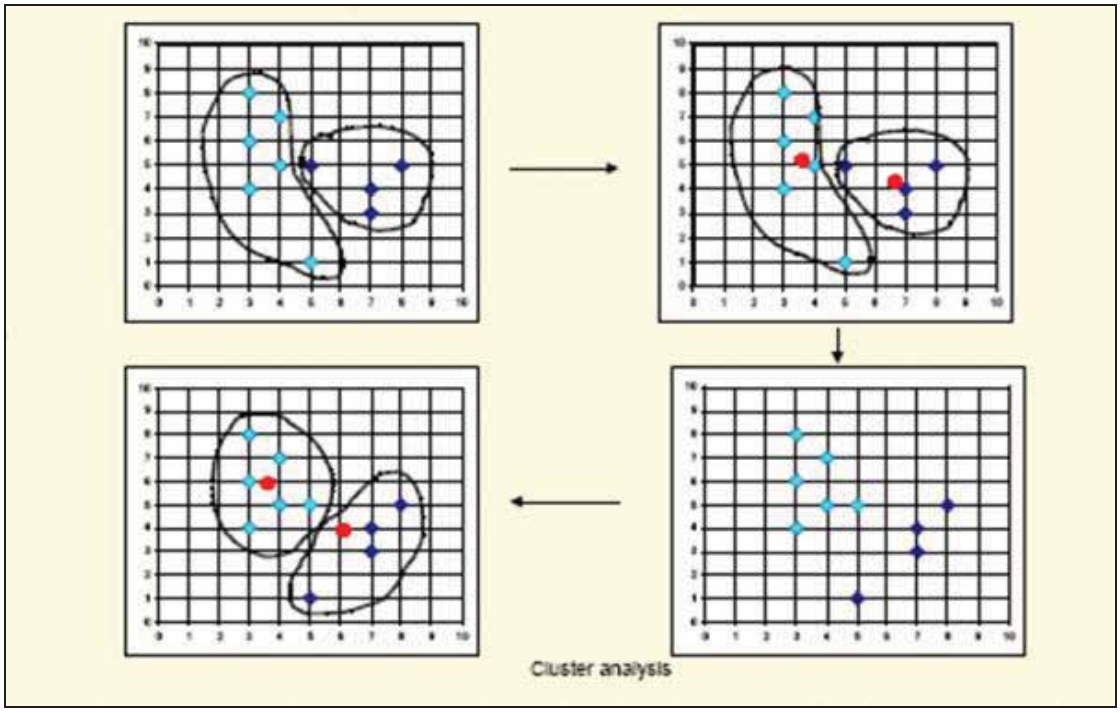
$$\text{Distance} = \sqrt{(x_1 - c_0)^2 + (y_1 - c_1)^2} \tag{4.5}$$

3.8.4 เมื่อได้กลุ่มข้อมูลใหม่แล้วทำการคำนวณหาจุดเซ็นทรอยด์ของกลุ่มนั้นใหม่ คำนวณจุดกึ่งกลาง (centroid) ของกลุ่ม โดยใช้ค่าเฉลี่ยเลขคณิต(mean) สูตรการหา centroid คือ

$$c = \left(\frac{x_1 + x_2 + \ldots + x_n}{m} \right), \left(\frac{y_1 + y_2 + \ldots + y_n}{m} \right) \tag{4.6}$$

3.8.5 ทำการคำนวณหาระยะทางแบบเดิมและทำการย้ายข้อมูลไปยังกลุ่มที่ทำให้ระยะห่างระหว่างข้อมูลกับจุดเซ็นทรอยด์มีค่าต่ำที่สุด ถ้าขั้นนี้ไม่มีการย้ายกลุ่มอีก แสดงว่ากลุ่มที่แบ่งได้นั้นเหมาะสมแล้ว แต่ถ้าในขั้นนี้ยังมีการย้ายกลุ่ม กลุ่มที่มีการย้ายเข้าย้ายออกจะต้องทำการคำนวณหาจุดเซ็นทรอยด์ใหม่ จนกว่าค่าของจุดเซ็นทรอยด์จะคงที่

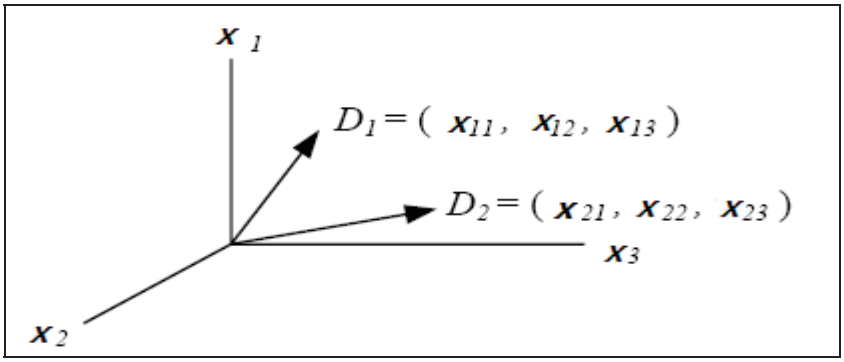
กำหนด k=2 สีฟ้าแทนกลุ่ม1 สีน้ำเงินแทนกลุ่ม 2 และสีแดงแทนค่าจุดเซ็นทรอยด์



ภาพที่ 32 แสดงการแบ่งกลุ่มข้อมูลโดยทฤษฎี K-means

ทฤษฎี K-Means นั้นยังสามารถนำมาประยุกต์ใช้กับข้อมูล n มิติได้ โดยสามารถนำข้อมูลที่ลดรูปแล้วจากทฤษฎี PCA มาจัดรูปให้อยู่ในรูป Model โดย

กำหนดให้ D คือ เว็บไซต์ และ X คือ จำนวน N มิติ ใด ๆ ของเว็บไซต์ D จากรูปจะแสดงกราฟกรณีลดรูปจาก PCA ให้เหลือ 3 มิติ ใด ๆ ซึ่งสามารถประยุกต์ใช้กับ N มิติใด ๆ ได้



ภาพที่ 33 แสดงการจัดรูป Model มิติที่ลดรูปแล้วด้วยเวกเตอร์

โดยสามารถคำนวณค่าระยะห่างระหว่างข้อมูลกับจุดกึ่งกลางได้จากสูตร

$$d(\mathrm{i,j}) = \sqrt{(x_{i1} - c_{j1})^2 + \ldots + (x_{ip} - c_{jp})^2} \quad (4.7)$$

เมื่อ d แทน Euclidean distance

c แทนจุดกึ่งกลาง

x แทนข้อมูล

i แทนลำดับของข้อมูล

j แทนลำดับของจุดกึ่งกลาง

p แทนมิติของข้อมูล

และ c จุดกึ่งกลางใหม่สามารถคำนวณได้จาก

$$\frac{x_{11} + x_{21} + \ldots + x_{i1}}{i} + \ldots + \frac{x_{11} + x_{21} + \ldots + x_{ip}}{i} \quad (4.8)$$

4. สร้างระบบป้องกันการเข้าถึงเว็บไซต์ไม่เหมาะสม

เมื่อได้กลุ่มคำที่ไม่เหมาะสมและจุดที่ใช้สำหรับการจำแนกเว็บไซต์ไม่เหมาะสมแล้ว ขั้นต่อไปเป็นการสร้างระบบป้องกันการเข้าถึงเว็บไซต์ไม่เหมาะสม ติดตั้งบนเครื่องแม่ข่ายที่มีโปรแกรม Squid-2.7 stable3 ทำหน้าที่เป็น Proxy cache server บนระบบปฏิบัติการ Ubuntu9.04

4.1 การเตรียมความพร้อมสำหรับระบบ

ระบบป้องกันการเข้าถึงเว็บไซต์ไม่เหมาะสม ใช้งานร่วมกับโปรแกรม squid โดยอ่านชื่อเว็บไซต์ที่ Client ร้องขอข้อมูล ซึ่งถูกบันทึกไว้ใน var/log/squid/access.log ระบบจะทำการร้องขอข้อมูล html code จากเว็บไซต์เพื่อนำมาจำแนกว่าเป็นเว็บไซต์ไม่เหมาะสม หรือเว็บไซต์ปกติ โดยตรวจสอบค่ามิติต่าง ๆ โดยนำค่ามิติต่าง ๆ เข้าสมการ PCA และทำการคำนวณค่าคงที่ของสมการเส้นตรงเพื่อใช้ตรวจสอบกับค่าคงที่ของสมการเส้นตรงที่คำนวณไว้แล้วหากเป็นเว็บไซต์ไม่เหมาะสมจะบันทึกชื่อเว็บไซต์ดังกล่าวลงใน blacklists ซึ่ง การเตรียมความพร้อมของระบบประกอบด้วย

4.1.1 การทำ Transparency

เพื่อเปลี่ยนทิศทางการร้องขอข้อมูลจาก web server ของโปรแกรมด้านเอกสารบนอินเทอร์เน็ต (browser) ซึ่งใช้ port 80 ให้ทำการร้องขอไปที่โปรแกรม squid ซึ่งใช้ port 8080 แทน โดยทำการเพิ่มข้อความต่อไปนี้ลงในไฟล์ /etc/rc.d/rc.local

iptables -t nat -F

iptables -t mangle -F

iptables -t filter -F

iptables -X

iptables -A FORWARD -j ACCEPT

iptables -t nat -A POSTROUTING -o eth0 -j MASQUERADE

iptables -t nat -A PREROUTING -i eth1 -p tcp --dport 80 -j REDIRECT --to-port 8080 และเพิ่มข้อความต่อไปนี้ลงในไฟล์ /etc/squid/squid.conf

http_port 8080 transparent

เพื่อเป็นการป้องกันไม่ให้เครื่องลูกข่ายเข้าใช้งานเว็บไซต์โดยไม่ผ่าน Proxy Cache Server

4.1.2 การปรับปรุงไฟล์ /etc/squid/squid.conf

ปรับปรุงไฟล์กำหนดสภาวะแวดล้อมการทำงานของSquid ทำการแก้ไขไฟล์ /etc/squid/squid.conf โดยการแทรกข้อความต่อไปนี้

acl ProxyFilter src 127.0.0.1/255.255.255.255

http_access allow ProxyFilter

acl lock1 url_regex '/home/ProxyFilter/blacklist1.txt'

http_access deny lock1

deny_info http://www.math26.com/thesis/warning1.html lock1

acl lock2 url_regex '/home/ProxyFilter/blacklist2.txt'

http_access deny lock2

deny_info http://www.math26.com/thesis/warning2.html lock2

เป็นการอนุญาตให้ระบบป้องกันการเข้าถึงเว็บไซต์ไม่เหมาะสมสามารถเข้าใช้ squid เพื่อร้องขอข้อมูล html code จากเว็บไซต์ได้ และสร้างระบบ Blacklists โดยใช้ไฟล์ blacklist1 และblacklist2 เป็นที่เก็บรายชื่อเว็บไซต์ไม่เหมาะสม ชื่อที่ถูกบันทึกลงในไฟล์นี้จะไม่สามารถเข้าใช้งานได้

4.2 การสร้างไฟล์ที่จำเป็นสำหรับการใช้งาน

4.2.1 ไฟล์ blacklist1.txt ไฟล์สำหรับเก็บรายชื่อเว็บไซต์ไม่เหมาะสม เกี่ยวกับความรุนแรง ยาเสพติด

4.2.2 ไฟล์ blacklist2.txt ไฟล์สำหรับเก็บรายชื่อเว็บไซต์ไม่เหมาะสม เกี่ยวกับคำลามกอนาจาร

4.2.3 ไฟล์ dict1 ไฟล์สำหรับเก็บกลุ่มคำความรุนแรง ยาเสพติดที่ใช้ในการจำแนกเว็บไซต์ไม่เหมาะสม โดย พิมพ์บรรทัดละ1 คำลงใน ไฟล์ dict1 โดยใช้กลุ่มคำความรุนแรงยาเสพติดที่ได้คัดเลือกแล้ว

4.2.4 ไฟล์ dict2 ไฟล์สำหรับเก็บกลุ่มคำลามกอนาจารที่ใช้ในการจำแนกเว็บไซต์ไม่เหมาะสม โดย พิมพ์บรรทัดละ1 คำลงใน ไฟล์ dict2 โดยใช้กลุ่มคำลามกอนาจารที่ได้คัดเลือกแล้ว

4.2.5 ไฟล์ dict3 ไฟล์เก็บกลุ่มคำเพิ่มเติม เกี่ยวกับ ความรุนแรง ยาเสพติด

4.2.6 ไฟล์ dict4 ไฟล์เก็บกลุ่มคำเพิ่มเติม เกี่ยวกับ ลามกอนาจาร

4.2.7 ไฟล์ whitelist.txt เป็นไฟล์สำหรับ เก็บรายชื่อ เว็บไซต์ปกติ

4.2.8 ไฟล์ proxy-filter.conf เป็นไฟล์ สำหรับกำหนดค่าสภาวะแวดล้อมในการทำงาน โดยประกอบไปด้วยข้อมูลดังต่อไปนี้

ACCESS_LOG=/var/log/squid/access.log

BLACK_LIST_1=/home/ProxyFilter/blacklist1.txt

BLACK_LIST_2=/home/ProxyFilter/blacklist2.txt

WHITE_LIST=/home/ProxyFilter/whitelist.txt

DICTIONARY_1=/home/ProxyFilter/dict1

DICTIONARY_2=/home/ProxyFilter/dict2

PROXYPORT=8080

LINE_CONSTANT _1= -35.9028

LINE_CONSTANT_2= -35.7921

ACCESS_LOG กำหนดไฟล์และตำแหน่งไฟล์ในการเก็บ Log ของเครื่อง Client

BLACK_LIST_1 กำหนดไฟล์และตำแหน่งไฟล์ที่เก็บรายชื่อเว็บไซต์ไม่เหมาะสม เกี่ยวกับ ความรุนแรง ยาเสพติด

BLACK_LIST_2 กำหนดไฟล์และตำแหน่งไฟล์ที่เก็บรายชื่อเว็บไซต์ไม่เหมาะสม เกี่ยวกับ ลามกอนาจาร

WHITE_LIST กำหนดไฟล์และตำแหน่งไฟล์ที่เก็บรายชื่อเว็บไซต์ปกติ

DICTIONARY_1 กำหนดไฟล์และตำแหน่งไฟล์ที่เก็บกลุ่มคำไม่เหมาะสมประเภทความรุนแรง ยาเสพติด

DICTIONARY_2 กำหนดไฟล์และตำแหน่งไฟล์ที่เก็บกลุ่มคำไม่เหมาะสมประเภทคำลามกอนาจาร

PROXYPORT กำหนดหมายเลข port ที่ใช้ติดต่อกับ Proxy Cache Server

LINE_CONSTANT_1 กำหนดค่าคงที่ที่ใช้ในการตรวจสอบเพื่อจำแนกเว็บไซต์ไม่เหมาะสมประเภทคำรุนแรง ยาเสพติด

LINE_CONSTANT_2 กำหนดค่าคงที่ที่ใช้ในการตรวจสอบเพื่อจำแนกเว็บไซต์ไม่เหมาะสมประเภทคำลามกอนาจาร

สรุประบบป้องกันการเข้าถึงเว็บไซต์ไม่เหมาะสมประกอบด้วยไฟล์ต่างๆดังนี้

/home/ProxyFilter/ProxyFilter ไฟล์เก็บ Config ของระบบ

/home/ProxyFilter/dict1 ไฟล์เก็บกลุ่มคำไม่เหมาะสม เกี่ยวกับคำรุนแรง ยาเสพติด

/home/ProxyFilter/dict2 ไฟล์เก็บกลุ่มคำไม่เหมาะสม เกี่ยวกับคำลามกอนาจาร

/home/ProxyFilter/dict3 ไฟล์เก็บกลุ่มคำไม่เหมาะสม เกี่ยวกับคำรุนแรง ยาเสพติดเพิ่มเติม

/home/ProxyFilter/dict4 ไฟล์เก็บกลุ่มคำไม่เหมาะสม เกี่ยวกับคำลามกอนาจารเพิ่มเติม

/home/ProxyFilter/whitelist ไฟล์เก็บรายชื่อเว็บไซต์ปกติ

/home/ProxyFilter/blacklist1 ไฟล์เก็บรายชื่อเว็บไซต์เกี่ยวกับ ความรุนแรง ยาเสพติด

/home/ProxyFilter/blacklist2 ไฟล์เก็บรายชื่อเว็บไซต์เกี่ยวกับ ลามกอนาจาร

/home/ProxyFilter/result.tmp ไฟล์ชั่วคราวเก็บ http return code

/home/ProxyFilter/httpclient.tmp ไฟล์ชั่วคราวเก็บ html code

/home/ProxyFilter/meta.tmp ไฟล์ชั่วคราวเก็บข้อมูล mata และ title

5. วิเคราะห์การจำแนกเว็บไซต์โดยใช้ทฤษฎี SVM (SVM : Support Vector Machine) เพื่อเปรียบเทียบประสิทธิภาพการคัดกรอง

ขั้นตอนการคัดกรองเว็บไซต์ด้วยวิธี SVM (SVM : Support Vector Machine)

5.1 กำหนด Feature ที่ใช้ในการทดสอบ

การกำหนด Feature ที่ใช้ในการทดสอบนั้นจะทำการกำหนดจากเนื้อหาในเว็บไซต์เพื่อใช้จำแนกความแตกต่าง ของข้อมูลในแต่ละเว็บไซต์โดย Feature ของ SVM ที่ใช้ในการทดสอบจะใช้ค่ามิติของ PCA เพื่อเปรียบเทียบผลการคัดกรองเว็บไซต์กับวิธี PCA ดังนั้นจึงทำการเลือกข้อมูลในการทดสอบชุดเดียวกัน

5.2 ขั้นตอนก่อนการประมวลผล (pre-processing)

เป็นขั้นตอนการเตรียมข้อมูล Feature ของเว็บไซต์โดยแบ่งการรวบรวม Feature เว็บไซต์ออกเป็น 2 ชุดได้แก่

5.2.1 Data Train คือเว็บไซต์ที่จะนำมาใช้ในการทดสอบให้ระบบ SVM สามารถเรียนรู้และสร้างโมเดลการคัดกรองเว็บไซต์ไม่เหมาะสมได้ โดยเว็บไซต์ที่เตรียมไว้นี้ได้ทำการแบ่งข้อมูล Data Train เป็น 2 กลุ่มย่อยคือ Data Train สำหรับการจำแนกเว็บไซต์รุนแรง ยาเสพติดโดยรวบรวมเว็บไซต์รุนแรง ยาเสพติด 200 เว็บไซต์ เว็บไซต์ปกติ 200 เว็บไซต์ และ Data Train สำหรับการจำแนกเว็บไซต์ลามกอนาจาร โดยรวบรวมเว็บไซต์ลามกอนาจาร 150 เว็บไซต์ และเว็บไซต์ปกติ 200 เว็บไซต์

5.2.2 Data Test คือเว็บไซต์ที่นำมาทดสอบกับโมเดลที่เราสร้างไว้แล้ว เพื่อทำการทดสอบว่าระบบสามารถคัดกรองเว็บไซต์ไม่เหมาะสมได้ถูกต้องมากน้อยเพียงใด โดยข้อมูล Data Test ได้แบ่งเป็น 2 กลุ่มย่อยคือData Test สำหรับทดสอบเว็บไซต์รุนแรง ยาเสพติดโดยรวบรวมเว็บไซต์รุนแรง ยาเสพติด 100 เว็บไซต์ เว็บไซต์ปกติ 100 เว็บไซต์ และ Data Train สำหรับการจำแนกเว็บไซต์ลามกอนาจาร โดยรวบรวมเว็บไซต์ลามกอนาจาร 100 เว็บไซต์ และเว็บไซต์ปกติ 100 เว็บไซต์

5.3 ขั้นตอนการประมวลผล (processing)

วิธีการคัดกรองเว็บไซต์ไม่เหมาะสมด้วยวิธี SVM นั้น จะใช้เครื่องมือ SVM^{light} V6.02 ซึ่งมีขั้นตอนในการสร้างระบบการคัดกรองดังนี้

5.3.1 นำข้อมูล Data Train ที่รวบรวมไว้มากำหนดให้เป็นรูปแบบเพื่อใช้ในการInput ข้อมูลเข้าโปรแกรม SVM^{light} โดยกำหนดให้หลักแรก คือ ค่า 1 แสดงถึงเว็บไซต์ไม่เหมาะสม และค่า -1 แสดงถึงเว็บไซต์ปกติ หลักต่อมาแสดงถึง Feature:ความถี่รวมของค่าFeature

5.3.2 นำข้อมูล training ที่เตรียมไว้ผ่านโปรแกรมเพื่อสร้างโมเดลโดยใช้คำสั่งในการสร้างโมเดลดังนี้ “svm_learn [option] example_file model_file” โดย

example_file คือ Data Train ที่มีการจัดรูปแบบแล้ว

Model_file คือ โมเดลเรียนรู้ของระบบซึ่งเป็น output ที่ได้

Option ที่ใช้ในการทดสอบมีดังนี้

Learning options:

-z{c,r,p} c หมายถึงการ Classification, r หมายถึงการ regression และ p หมายถึงการ preference ranking ซึ่งค่าที่เราใช้คือค่า c การ Classification

Kernel Options:

-t int คือ ชนิดของ kernel function

0: linear

1: polynomial

2: radial basis function (rbf)

3: sigmoid

-d int ค่าพารามิเตอร์ d ใน polynomial kernel

-g float ค่าพารามิเตอร์ gamma ใน rbf kernel

-s float ค่าพารามิเตอร์ s ใน sigmoid kernel

นำข้อมูล Data Train สำหรับทดสอบเว็บไซต์รุนแรงและทดสอบเว็บไซต์ลามกอนาจารมาทำการสร้างโมเดลและเลือก option สำหรับการทดสอบ

5.3.3 นำข้อมูล Data Test มาทดสอบกับ model การเรียนรู้ของระบบ โดยใช้คำสั่งในการทดสอบระบบดังนี้ "svm_classify example_file model_file output_file" โดย

example_file คือ Data Test ที่มีการจัดรูปแบบแล้ว

Model_file คือ โมเดลเรียนรู้ของระบบจากการสร้างไว้ในข้อ 5.3.2

Output_file คือ output ข้อมูลที่ได้จากการทดสอบข้อมูล Data Test กับระบบ model การเรียนรู้ ถ้าค่า output ที่แสดงได้เป็น + แสดงว่าเป็นข้อมูลกลุ่มเดียวกับ +1 คือเว็บไซต์ไม่เหมาะสม ถ้าค่า output ที่แสดงได้เป็น – แสดงว่าเป็นข้อมูลกลุ่มเดียวกับ –1 คือเว็บไซต์ปกติ

5.3.4 ทำการทดสอบแบ่งกลุ่มเว็บไซต์โดยใช้ทฤษฎี kernel functions กำหนดให้พารามิเตอร์ในแต่ละ kernel function มีค่าเป็นดังนี้

Polynomial กำหนดให้ d = 2, 3

RBF กำหนดให้ g = 0.01, 0.1, 0.5 และ 1

Sigmoid กำหนดให้ s = 0.25, 0.5, 1 และ 2

5.4 ผลการทดสอบการแบ่งกลุ่มเว็บไซต์โดยใช้ทฤษฎี SVM (SVM : Support Vector Machine)

เป็นขั้นตอนของการนำผลเว็บไซต์ที่ทำการแบ่งประเภทเว็บไซต์แล้วมาประเมินประสิทธิภาพ โดยข้อมูลทดสอบการคัดกรองเว็บไซต์เป็นข้อมูลชุดเดียวกับการทดสอบ PCA

1. ทดสอบเว็บไซต์ความรุนแรง ยาเสพติด แบ่งเป็นเว็บไซต์ปกติ 100 เว็บไซต์ และเว็บไซต์ความรุนแรง ยาเสพติด 100 เว็บไซต์

2. ทดสอบเว็บไซต์ลามกอนาจาร แบ่งเป็นเว็บไซต์ปกติ 100 เว็บไซต์ และเว็บไซต์ลามกอนาจาร 100 เว็บไซต์

ทำการตรวจสอบดูว่าเว็บไซต์ที่แบ่งประเภทนั้นมีค่าเป็นอย่างไร โดยทำการประเมินประสิทธิภาพค่าต่าง ๆ ตามตารางดังนี้

ตารางที่ 18 แสดงประสิทธิภาพการคัดกรองเว็บไซต์ด้วยวิธี SVM แบบ linear

ประเภทเว็บไซต์	ค่าความถูกต้อง Accuracy	ค่าความแม่นยำ Precision	ค่าความระลึก Recall
เว็บไซต์ความรุนแรง ยาเสพติด	89	90.63	87
เว็บไซต์ลามกอนาจาร	87	100	74

การคัดกรองเว็บไซต์ด้วยวิธี SVM แบบ linear พบว่าประเภทเว็บไซต์ความรุนแรง ยาเสพติด มีค่าความถูกต้องคิดเป็นร้อยละ89 และประเภทเว็บไซต์ลามกอนาจาร มีค่าความถูกต้องคิดเป็นร้อยละ 87 ใช้เวลาในการสร้างแต่ละ Model 1.35 วินาที

ตารางที่ 19 แสดงประสิทธิภาพการคัดกรองเว็บไซต์ด้วยวิธี SVM แบบ Polynomial kernel

ประเภทเว็บไซต์	พารามิเตอร์ของ Polynomial Kernel	ค่าความถูกต้อง Accuracy	ค่าความแม่นยำ Precision	ค่าความระลึก Recall
เว็บไซต์ความรุนแรง ยาเสพติด	2	66	59.88	97
	3	50	0	0
เว็บไซต์ลามกอนาจาร	2	71.5	66.67	86
	3	91	97.67	84

การคัดกรองเว็บไซต์ด้วยวิธี SVM แบบ Polynomial kernel นั้นได้ทำการทดสอบโดยใช้พารามิเตอร์ 2 ค่า ได้แก่ พารามิเตอร์ 2 พบว่าประเภทเว็บไซต์ความรุนแรงยาเสพติด มีค่าความถูกต้องคิดเป็นร้อยละความถูกต้อง 66 และประเภทเว็บไซต์ลามกอนาจาร มีค่าความถูกต้องคิด

เป็นร้อยละความถูกต้อง 50 พารามิเตอร์ 3 พบว่าประเภทเว็บไซต์ความรุนแรงยาเสพติด มีค่าความถูกต้องคิดเป็นร้อยละ 71.5 และประเภทเว็บไซต์ลามกอนาจาร มีค่าความถูกต้องคิดเป็นร้อยละ 91

สรุปเว็บไซต์ความรุนแรง ยาเสพติด คัดกรองด้วยพารามิเตอร์ 2 มีประสิทธิภาพมากกว่า ส่วนเว็บไซต์ลามกอนาจารคัดกรองด้วยพารามิเตอร์ 3 มีประสิทธิภาพมากกว่า กรณี Polynomial Degree 2 ใช้เวลาในการสร้าง Model 1.5 วินาที กรณี Polynomial Degree 3 ใช้เวลาในการสร้าง Model 1 ชั่วโมง

ตารางที่ 20 แสดงประสิทธิภาพการคัดกรองเว็บไซต์ด้วยวิธี SVM แบบ RBF kernel

ประเภทเว็บไซต์	พารามิเตอร์ของ RBF Kernel	ค่าความถูกต้อง Accuracy	ค่าความแม่นยำ Precision	ค่าความระลึก Recall
เว็บไซต์ความรุนแรงยาเสพติด	0.01	80	100	60
	0.1	74.5	100	49
	0.5	73.5	100	47
	1	73.5	100	47
เว็บไซต์ลามกอนาจาร	0.01	79	100	58
	0.1	57	100	14
	0.5	52	100	4
	1	51	100	2

การคัดกรองเว็บไซต์ด้วยวิธี SVM แบบ RBF kernel นั้นได้ทำการทดสอบโดยใช้ค่าพารามิเตอร์ดังนี้ 0.01 0.1 0.5 และ 1 ผลการทดสอบพบว่าพารามิเตอร์ 0.01 ให้ผลความถูกต้องมากที่สุด สามารถการคัดกรองเว็บไซต์ความรุนแรง ยาเสพติดได้ถูกต้องคิดเป็นร้อยละ 80 และเว็บไซต์ลามกอนาจารให้ผลความถูกต้องคิดเป็นร้อยละ 79 ใช้เวลาในการสร้างแต่ละ Model 1.2 วินาที

ตารางที่ 21 แสดงประสิทธิภาพการคัดกรองเว็บไซต์ด้วยวิธี SVM แบบ Sigmoid kernel

ประเภทเว็บไซต์	พารามิเตอร์ของ Sigmoid Kernel	ค่าความถูกต้อง Accuracy	ค่าความแม่นยำ Precision	ค่าความระลึก Recall
เว็บไซต์ความรุนแรงยาเสพติด	0.25	50	50	100
	0.5	50	50	100
	1	50	50	100
	2	50	50	100
เว็บไซต์ลามกอนาจาร	0.25	50	50	100
	0.5	50	50	100
	1	50	50	100
	2	50	0	0

การคัดกรองเว็บไซต์ด้วยวิธี SVM แบบ Sigmiod kernael นั้นได้ทำการทดสอบโดยใช้ค่าพารามิเตอร์ดังนี้ 0.25 0.5 1 2 ผลการทดสอบการคัดกรองเว็บไซต์ความรุนแรง ยาเสพติด และเว็บไซต์ลามกอนาจารได้ค่าความถูกต้องเท่ากันหมด ซึ่งไม่สามารถคัดกรองเว็บไซต์ได้ด้วยวิธีนี้ ใช้เวลาในการสร้างแต่ละ Model 4.2 ชั่วโมง

5.5 ทำการเปรียบเทียบผลการทดสอบการคัดกรองเว็บไซต์ด้วยวิธี PCA (PCA : Principal Component Analysis) กับวิธี SVM (SVM : Support Vector Machine)

วิธีการแบ่งกลุ่มข้อมูลโดยใช้วิธี SVM นั้น ส่วนมากการแบ่งกลุ่มข้อมูลที่เป็นเอกสาร หรือ ข้อความจะใช้วิธี linear หรือ วิธี Polynomial ส่วนวิธีการแบ่งกลุ่มข้อมูลที่เป็นภาพนั้นจะใช้วิธี RBF (Researchers SVM, 2010) จากทฤษฎีนี้จึงพิจารณาการแบ่งกลุ่มข้อมูลเว็บไซต์ด้วยวิธี linear และ วิธี Polynomial ซึ่งจากผลการทดลองจะเห็นได้ว่าวิธีการคัดกรองเว็บไซต์ประเภทความรุนแรง ยาเสพติดใช้ Kernel Function แบบ liner ให้ค่าถูกต้องมากที่สุด และวิธีการคัดกรองเว็บไซต์ประเภทลามกอนาจารใช้ Kernel Function แบบ Polynomial Degree 3 เป็นวิธีการคัดกรองเว็บไซต์ที่ให้ค่าถูกต้องมากที่สุดซึ่งจะนำมาเปรียบเทียบกับวิธี PCA โดยทำการทดสอบกับเว็บไซต์ความรุนแรง ยาเสพติด 100 เว็บไซต์ เว็บไซต์ปกติประเภทความรุนแรง ยาเสพติด 100 เว็บไซต์ และเว็บไซต์ลามกอนาจาร 100 เว็บไซต์ เว็บไซต์ปกติประเภทลามกอนาจาร 100 เว็บไซต์

ตารางที่ 22 แสดงผลการเปรียบเทียบประสิทธิภาพการคัดกรองเว็บไซต์ความรุนแรง ยาเสพติดด้วยวิธี PCA กับวิธี SVM

ประเภทเว็บไซต์	จำนวนเว็บ	คัดกรองเว็บไซต์ด้วยวิธี PCA				คัดกรองเว็บไซต์ด้วยวิธี SVM			
		จำแนกถูก		จำแนกผิด		จำแนกถูก		จำแนกผิด	
		จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ
เว็บไซต์ความรุนแรงยาเสพติด	100	82	82	18	18	87	87	13	13
เว็บไซต์ปกติประเภทรุนแรงยาเสพติด	100	97	97	3	3	91	91	9	9
เฉลี่ยรวม	200	179	89.5	21	10.5	178	89	22	11

ตารางที่ 23 แสดงผลการเปรียบเทียบประสิทธิภาพการคัดกรองเว็บไซต์ลามกอนาจารด้วยวิธี PCA กับวิธี SVM

ประเภทเว็บไซต์	จำนวนเว็บ	คัดกรองเว็บไซต์ด้วยวิธี PCA				คัดกรองเว็บไซต์ด้วยวิธี SVM			
		จำแนกถูก		จำแนกผิด		จำแนกถูก		จำแนกผิด	
		จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ
เว็บไซต์ความลามกอนาจาร	100	89	89	11	11	84	84	16	16
เว็บไซต์ปกติประเภทลามกอนาจาร	100	100	100	0	0	98	98	2	2
เฉลี่ยรวม	200	189	94.5	11	5.5	182	91	18	9

จากผลการทดลองสรุปได้ว่ากรณีเว็บไซต์ความรุนแรง ยาเสพติดวิธี PCA สามารถคัดกรองเว็บไซต์ได้ถูกต้องทั้งหมดคิดเป็นร้อยละ 89.5 เปอร์เซ็นต์ ส่วนวิธี SVM นั้น สามารถคัดกรองเว็บไซต์ได้ถูกต้องทั้งหมดคิดเป็นร้อยละ 89 เปอร์เซ็นต์ ส่วนกรณีเว็บไซต์ลามกอนาจาร วิธี PCA

สามารถคัดกรองเว็บไซต์ได้ถูกต้องทั้งหมดคิดเป็นร้อยละ 94.5 เปอร์เซ็นต์ ส่วนวิธี SVM นั้น สามารถคัดกรองเว็บไซต์ได้ถูกต้องทั้งหมดคิดเป็นร้อยละ 91 เปอร์เซ็นต์

ดังนั้นสรุปได้ว่าการคัดกรองเว็บไซต์กลุ่มคำรุนแรง ยาเสพติด และกลุ่มเว็บไซต์ลามกอนาจารนั้น การคัดกรองด้วยวิธี PCA นั้นมีความถูกต้องมากกว่าวิธีการคัดกรองด้วย SVM

6. วิเคราะห์การเพิ่มกลุ่มคำไม่เหมาะสมแบบอัตโนมัติ

กลุ่มคำบางกลุ่มคำที่ผู้ใช้ระบบไม่แน่ใจว่าเป็นกลุ่มคำที่มีอำนาจการแจกแจงที่เหมาะสมต่อการนำไปใช้ในการ Block เว็บไซต์หรือไม่ โดยผู้ใช้สามารถเพิ่มกลุ่มคำที่ไม่แน่ใจในระบบเพื่อให้ระบบช่วยตรวจสอบให้

6.1 กรณีประเภทกลุ่มคำรุนแรง ยาเสพติด มีวิธีการตรวจสอบดังนี้

6.1.1 การเพิ่มกลุ่มคำที่ไม่แน่ใจว่าเป็นคำไม่เหมาะสมประเภทคำรุนแรง โดยทำการเพิ่มเข้าไปในระบบ 50 คำดังนี้

ฆ่าฟัน ทำลายล้าง ย่ำยี ยื้อยุด พุ่งชน ดักทำร้าย ซ้อมน่วม ฉุดกระชาก สหบาทา อยากยา เลือดคลั่ง ปลิดชีพ ลงแดง ยาบ้า มีดพก ไม้หวด ขู่ฆ่า รุมสกัม อาการสาหัส รุมซ้อม มอมยา ซุ่มสังหาร ดับดิ้น ยกพวก สะเทือนขวัญ รุมยำ หวาดเสียว รุมตื้บ นองเลือด จิกผม วางมวย ตบตี ระดมยิง ปลอกกระสุนปืน ปืนปากกา เป้าสังหาร ตะลุมบอล ลั่นไก หวีดร้อง พยายามฆ่า สาดน้ำกรด รุมทำร้าย ยาม้า รุมตบ สำเร็จความใคร่ นักฆ่า คุ้มคลั่ง สลบเหมือด ชำระแค้น ป่าเถื่อน

6.1.2 ระบบจะทำการนับความถี่สะสม ของคำแต่ละคำ ในกรณีที่ผู้ใช้มีการเรียกใช้งานเว็บไซต์ที่ไม่เหมาะสมประเภทความรุนแรง ยาเสพติด

6.1.3 คำนวณจุดความถี่ที่เหมาะสมในการเพิ่มกลุ่มคำไม่แน่ใจให้เป็นกลุ่มคำที่ไม่เหมาะสมประเภทกลุ่มคำรุนแรง ยาเสพติดเพื่อใช้ Block เว็บไซต์คือ

จุดความถี่ที่เหมาะสม = (ค่าอำนาจการแจกแจงที่มากที่สุด+ค่าอำนาจการแจกแจงที่น้อยที่สุด) / 2 จากข้อมูลการคำนวณหาอำนาจการแจกแจงกลุ่มคำรุนแรง ยาเสพติด

ค่าอำนาจการแจกแจงที่มากที่สุดคือ แทง มีค่า 24.50

ค่าอำนาจการแจกแจงที่น้อยที่สุดคือ รุมโทรม บุกยิง มีค่า 10.25

จุดความถี่ที่เหมาะสม = (24.50 +10.25) / 2 = 17.375

ดังนั้นจุดความถี่ที่เหมาะสมในการเพิ่มกลุ่มคำที่ไม่แน่ใจประเภทกลุ่มคำรุนแรง ยาเสพติดคือ 17 คำ หากนับความถี่สะสม ใน 50 คำนี้ เป็นจำนวน 200 รอบ ถ้าคำใดมีความถี่สะสมเท่ากับ 17 คำ ถือว่าคำนั้นมีอำนาจการแจกแจงเหมาะสม ในการคัดกรองเว็บไซต์

6. 2 กรณีประเภทกลุ่มคำลามกอนาจาร มีวิธีการตรวจสอบดังนี้

6.2.1 การเพิ่มกลุ่มคำที่ไม่แน่ใจว่าเป็นคำไม่เหมาะสมประเภทคำลามกอนาจาร โดยทำการเพิ่มเข้าไปในระบบ 50 คำดังนี้

คำอนาจารที่63 คำอนาจารที่ 64 คำอนาจารที่65 คำอนาจารที่66 คำอนาจารที่67 คำอนาจารที่68 คำอนาจารที่69 คำอนาจารที่70 คำอนาจารที่71 คำอนาจารที่72 คำอนาจารที่73 คำอนาจารที่74 คำอนาจารที่75 คำอนาจารที่76 คำอนาจารที่77 คำอนาจารที่78 คำอนาจารที่79 คำอนาจารที่80 คำอนาจารที่81 คำอนาจารที่82 คำอนาจารที่83 คำอนาจารที่84 คำอนาจารที่85 คำอนาจารที่86 คำอนาจารที่87 คำอนาจารที่88 คำอนาจารที่89 คำอนาจารที่90 คำอนาจารที่91 คำอนาจารที่92 คำอนาจารที่93 คำอนาจารที่94 คำอนาจารที่95 คำอนาจารที่96 คำอนาจารที่97 คำอนาจารที่98 คำอนาจารที่99 คำอนาจารที่100 คำอนาจารที่101 คำอนาจารที่102 คำอนาจารที่103 คำอนาจารที่104 คำอนาจารที่105 คำอนาจารที่106 คำอนาจารที่107 คำอนาจารที่108 คำอนาจารที่109 คำอนาจารที่110 คำอนาจารที่111 คำอนาจารที่112

6.2.2 ระบบจะทำการนับความถี่สะสมของคำแต่ละคำ ในกรณีที่ผู้ใช้มีการเรียกใช้งานเว็บไซต์ ที่ไม่เหมาะสมประเภทลามกอนาจาร

6.2.3 คำนวณจุดความถี่ที่เหมาะสมในการเพิ่มกลุ่มคำไม่แน่ใจให้เป็นกลุ่มคำที่ไม่เหมาะสม ประเภทกลุ่มคำลามกอนาจารเพื่อใช้ Block เว็บไซต์คือ

จุดความถี่ที่เหมาะสม = (ค่าอำนาจการแจกแจงที่มากที่สุด + ค่าอำนาจการแจกแจงที่น้อยที่สุด) / 2

จากข้อมูลการคำนวณหาอำนาจการแจกแจงกลุ่มคำลามกอนาจาร ค่าอำนาจการแจกแจงที่มากที่สุดคือ คำอนาจารที่14 มีค่า 255 ซึ่งมีค่าอำนาจการแจกแจงห่างจากค่าอื่นมากเกินไปดังนั้นจึงทำการเลือกค่าอำนาจการแจกแจงที่มากที่สุดรองลงมาคือ คำอนาจารที่ 27 มีค่า 102

ค่าอำนาจการแจกแจงที่น้อยที่สุดคือ คำอนาจารที่17 มีค่า 43.67

จุดความถี่ที่เหมาะสม = (102 + 43.67) / 2 = 72.835

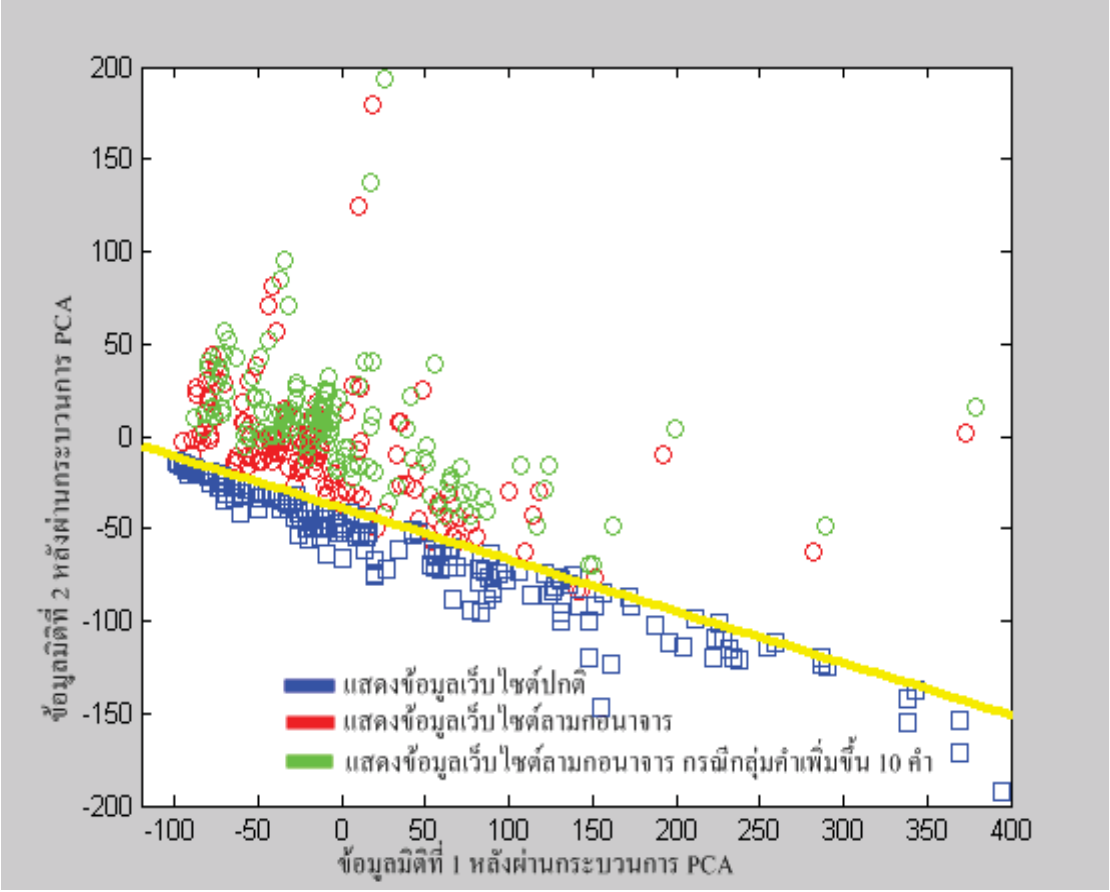
ดังนั้นจุดความถี่ที่เหมาะสมในการเพิ่มกลุ่มคำที่ไม่แน่ใจประเภทกลุ่มคำลามกอนาจารคือ 72 คำ หากนับความถี่สะสม ใน 50 คำนี้ เป็นจำนวน 150 รอบ ถ้าคำใดมีความถี่สะสมเท่ากับ 72 คำ ถือว่าคำนั้นมีอำนาจการแจกแจงเหมาะสม ในการคัดกรองเว็บไซต์

การเพิ่มกลุ่มคำอัตโนมัติของระบบมีผลทำให้การคัดกรองข้อมูลมีประสิทธิภาพมากขึ้น ทำการทดสอบเพิ่มกลุ่มคำใน (Dict1, Dict2) โดยเพิ่มกลุ่มคำให้มากขึ้นตามลำดับเมื่อมีการเพิ่มกลุ่มคำใน Dict1, Dict2 แล้ว จะทำให้ระบบคัดกรองเว็บไซต์ที่ไม่เหมาะสมมีประสิทธิภาพมากขึ้น

จากทฤษฎี PCA ซึ่งจำนวนคำไม่เหมาะสมใน POOL และ จำนวนคำที่ไม่เหมาะสมใน BODY TAG ซึ่งเป็นมิติที่เป็นปัจจัยต่อการจำแนกเว็บไซต์

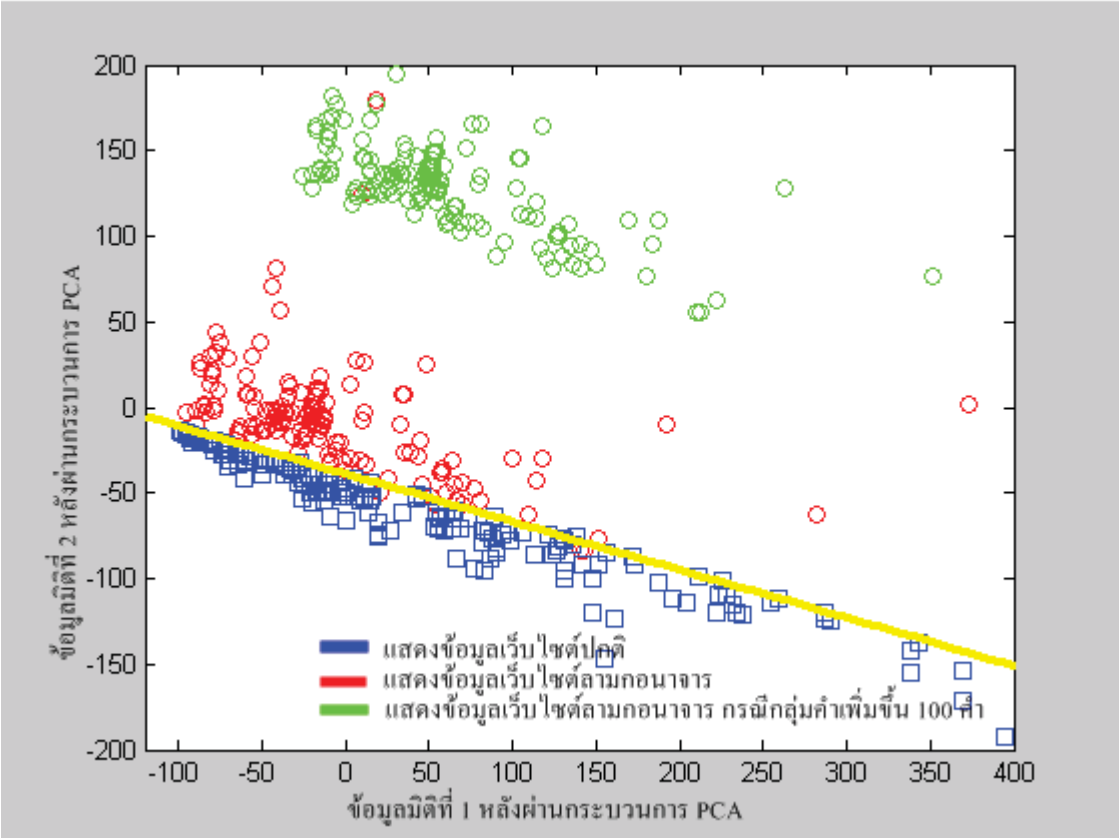
ดังนั้นจะทำการทดสอบโดยเพิ่มกลุ่มคำที่ไม่เหมาะสมเพิ่มขึ้นเรื่อย ๆ ตามลำดับ และนำมิติจำนวนคำไม่เหมาะสมใน POOL และ จำนวนคำที่ไม่เหมาะสมใน BODY TAG ไปทำการหาสมการ FinalData ใหม่โดยให้มิติอื่น ๆคงที่

ทดสอบกลุ่มตัวอย่างที่ 1 เพิ่มกลุ่มคำลามกอนาจารใน Dict2 อย่างละ 10 คำ



ภาพที่ 34 แสดงผลการคัดกรองเว็บไซต์กรณีกลุ่มคำเพิ่มขึ้น 10 คำ

จากกราฟจะเห็นได้ว่าเมื่อเพิ่มกลุ่มคำลามกจะทำให้ระบบคัดกรองเว็บไซต์มีประสิทธิภาพมากขึ้น ทดสอบกลุ่มตัวอย่างที่ 1 เพิ่มกลุ่มคำลามกอนาจารใน Dict2 อย่างละ 100 คำ



ภาพที่ 35 แสดงผลการคัดกรองเว็บไซต์กรณีกลุ่มคำเพิ่มขึ้น 100 คำ

จากกราฟจะเห็นได้ว่าเมื่อเพิ่มกลุ่มคำลามกจะทำให้ระบบคัดกรองเว็บไซต์มีประสิทธิภาพมากขึ้น

7. การทดสอบประสิทธิภาพการทำงาน

การทดสอบประสิทธิภาพการจำแนกเว็บไซต์ไม่เหมาะสมออกจากเว็บไซต์ปกติ แบ่งการทดสอบออกเป็น 2 ส่วนคือ ประสิทธิภาพด้านความถูกต้อง, ประสิทธิภาพด้านความเร็วและประสิทธิภาพในการรองรับเว็บ 2.0

7.1 ประสิทธิภาพด้านความถูกต้อง

การทดสอบประสิทธิภาพ ของระบบป้องกันการเข้าถึงเว็บไซต์ไม่เหมาะสมในด้านความถูกต้อง ใช้เว็บไซต์ในการทดสอบจำนวน 431 เว็บไซต์ ซึ่งได้จากบันทึกการใช้งานของเครื่องลูกข่ายที่ถูกบันทึกไว้ในเครื่องแม่ โดยใช้ไฟล์หน้าโฮมเพจไฟล์เดียวในการจำแนกผลที่ได้ จากการจำแนกของระบบ แสดงดังตาราง

ตารางที่ 24 แสดงประสิทธิภาพด้านความถูกต้องของการจำแนกเว็บไซต์

ประเภทเว็บไซต์	จำนวนเว็บไซต์	จำแนกถูก		จำแนกผิด	
		จำนวนเว็บ	ร้อยละ	จำนวนเว็บ	ร้อยละ
เว็บไซต์ความรุนแรงและยาเสพติด	156	129	82.69	27	17.31
เว็บไซต์ความลามกอนาจาร	155	131	84.52	24	15.48
เว็บไซต์ปกติ	120	105	87.5	15	12.5
เฉลี่ยรวม	431	365	84.69	66	15.31

จากตารางการวัดประสิทธิภาพด้านความถูกต้องของการจำแนกเว็บไซต์จากจำนวนเว็บไซต์ทั้งหมด 431 เว็บไซต์ สามารถจำแนกได้ถูกต้อง จำนวน 365 เว็บไซต์ คิดเป็นร้อยละ 84.69 ของจำนวนเว็บไซต์ทั้งหมด แบ่งเป็นสามารถจำแนกเว็บไซต์ความรุนแรง ยาเสพติดได้ถูกต้อง 129 เว็บไซต์ คิดเป็นร้อยละ 82.69 จำแนกเว็บไซต์ลามกอนาจารได้ถูกต้อง 131 เว็บไซต์ คิดเป็นร้อยละ 84.52 จำแนกเว็บไซต์ปกติได้ถูกต้อง 105 เว็บไซต์ คิดเป็นร้อยละ 87.5 และจำแนกเว็บไซต์ผิดพลาดทั้งหมดจำนวน 66 เว็บไซต์ คิดเป็นร้อยละ 15.31 ของจำนวนเว็บไซต์ทั้งหมด แบ่งเป็นจำแนกเว็บไซต์ความรุนแรง ยาเสพติดผิดพลาด 27 เว็บไซต์ คิดเป็นร้อยละ 17.31 จำแนกเว็บไซต์ลามกอนาจารผิดพลาด 24 เว็บไซต์ คิดเป็นร้อยละ 15.48 จำแนกเว็บไซต์ปกติผิดพลาด 15 เว็บไซต์ คิดเป็นร้อยละ 12.5 สำหรับความผิดพลาดที่ไม่สามารถจำแนกเว็บไซต์ไม่เหมาะสมได้มีสาเหตุมาจาก บางเว็บไซต์มีความถี่ของกลุ่มคำ น้อยกว่า เกณฑ์ที่กำหนดไว้ บางเว็บไซต์ใส่เนื้อหาไว้ในเว็บเพจอื่นๆ ที่ไม่ใช่หน้าโฮมเพจ บางเว็บไซต์มีการเปลี่ยนแปลงอยู่ตลอดเวลา บางเว็บไซต์ต้องสมัครสมาชิกก่อน จึงจะเห็นเนื้อหาด้านใน และความผิดพลาดที่เกิดมาจากการจำแนกเว็บไซต์ปกติผิดพลาด มีสาเหตุมาจาก เว็บไซต์ปกติมีกลุ่มคำที่ไม่เหมาะสม มากกว่าเกณฑ์ที่กำหนดไว้ จึงทำให้เกิดความผิดพลาด ทั้งๆ ที่เป็นเว็บไซต์ปกติ

7.2 ประสิทธิภาพด้านความเร็ว

ปัจจัยที่มีผลกระทบต่อประสิทธิภาพด้านความเร็วของระบบตรวจสอบการเข้าถึงเว็บไซต์ที่ไม่เหมาะสม มีหลายปัจจัย เช่น ความเร็วในการเชื่อมต่ออินเตอร์เน็ต จำนวนกลุ่มคำที่

ไม่เหมาะสม การคำนวณความถี่ของจำนวนกลุ่มคำเพิ่มเติม แต่ปัจจัยที่มีผลกระทบต่อระบบมากที่สุดคือ ปริมาณผู้ใช้งานอินเตอร์เน็ต ซึ่งเป็นปัจจัยหนึ่งที่ใช้ในการทดสอบ

จากผลการทดลองนำเว็บไซต์ 50 เว็บไซต์มาทดสอบกับปริมาณผู้ใช้งานอินเทอร์เน็ตที่เพิ่มขึ้นเรื่อย ๆ พบว่า เมื่อผู้ใช้งานอินเทอร์เน็ตเพิ่มขึ้นเวลาที่ใช้ในการประมวลผลของโปรแกรมเพิ่มขึ้น

ตารางที่ 25 แสดงถึงเวลาที่ใช้ในการ Block เว็บไซต์(วินาที) ในช่วงของผู้ใช้งานอินเทอร์เน็ตที่เพิ่มขึ้นเรื่อย

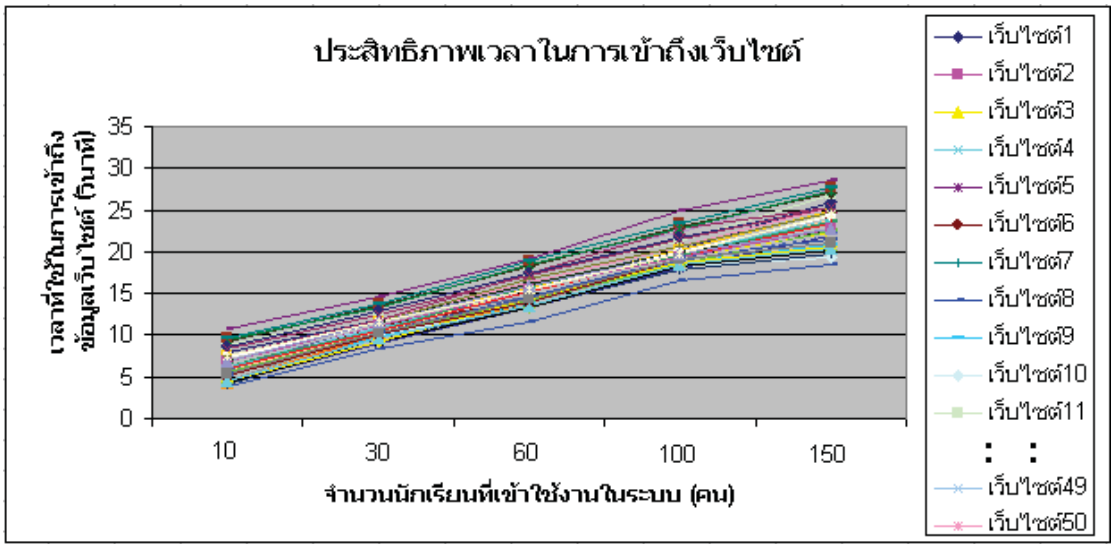
เวลาที่ใช้ในการเข้าถึงข้อมูลเว็บไซต์	**จำนวนนักเรียนที่เข้าใช้งานในระบบ**				
	10	**30**	**60**	**100**	**150**
http://atcloud.com/stories/15199	07.9	11.8	15.9	20.2	25.9
http://webboard.yenta4.com/topic/251441	05.2	10.3	14.6	19.4	23.6
http://talk.mthai.com/topic/44394	05.1	10.1	14.5	19.1	22.4
http://www.yorkza.com/content/7979/	06.5	10.9	15.9	19.7	23.9
http://atcloud.com/stories/15199	04.5	09.1	13.4	18.3	20.4
http://www.sodamag.net	05.9	11.2	14.2	19.2	23.9
www.Exteen.com	04.5	10.7	13.6	18.4	20.6
www.Ohozaa.com	03.9	08.4	11.5	16.6	18.4
www.Meemodel.com	04.5	09.9	13.9	18.7	20.9
http://www.madoo.com	04.2	10.1	13.4	18.3	19.4
http://regist53.blogspot.com/2009/09/3_28.html	08.9	12.8	17.5	21.2	26.9
http://news.mthai.com/general-news/52345.html	06.2	10.3	15.6	19.4	23.4
http://tnews.teenee.com/crime/583.html	06.1	11.1	15.5	19.1	22.1
http://board.postjung.com/423124.html	07.5	11.9	16.4	20.7	25.3
http://webboard.yenta4.com/topic/251441	09.5	13.5	18.4	22.3	27.4
http://bbs.soizaa.com/archiver/tid-8228.html	05.7	10.2	14.3	19.1	23.9
http://www.pitakthai.com/social/217.html	05.2	10.3	14.2	18.8	22.1

ตารางที่ 25 (ต่อ)

เวลาที่ใช้ในการเข้าถึงข้อมูลเว็บไซต์	จำนวนนักเรียนที่เข้าใช้งานในระบบ				
	10	**30**	**60**	**100**	**150**
http://www.ohthai.net/10732.html	06.3	10.8	15.6	19.8	23.6
http://www.ryt9.com/s/bmnd/730195/	04.4	09.2	13.3	18.2	20.1
http://atcloud.com/stories/45137	07.5	11.6	15.7	20.1	24.9
www.yenta4.com	04.2	10.4	13.5	18.4	20.4
www.siamzone.com	04.3	10.1	14.3	18.5	20.5
www.tlcthai.com	05.8	11.5	14.5	19.3	23.7
www.tarad.com	04.5	10.7	13.9	18.6	20.7
www.siamha.com	04.1	09.6	13.3	17.8	19.7
http://ss.comparenotebook.info/goURL/180.html	06.5	10.5	15.8	19.6	23.7
http://www.over18x.com/Breakspells-vol-3.xxx	06.8	11.7	16.1	19.9	22.6
http://www8.mobileacce.info/last/13152.html	09.2	13.6	18.3	22.9	27.1
http://avzone.wordpress.com/2009/01/	09.7	13.8	18.8	23.3	27.8
http://www.thaizexstory.com/home/story/128	07.5	11.9	17.4	22.7	25.3
http://talk.mthai.com/topic/20383	04.2	09.1	13.4	18.1	21.8
http://www.ryt9.com/s/bmnd/712912	04.4	09.2	13.3	18.2	20.1
http://news.mthai.com/general-news/52345.html	07.3	11.3	15.5	19.9	24.2
http://webboard.yenta4.com/topic/251441	06.1	10.6	15.3	19.4	23.5
http://www.pitakthai.com/crime/3487.html	05.5	10.2	14.1	18.5	22.6
www.365jukebox.com	05.4	10.4	14.5	19.4	21.4
www.zuzaa.com	04.5	09.3	14.4	18.7	20.6
http://www.jikgo.com	06.8	11.1	15.5	19.3	22.7
http://www.siamdara.com	04.5	09.7	13.5	18.4	20.5
www.zubzip.com	05.1	10.1	13.9	19.2	21.1
http://www3.lyricscom.info/lasted_stories/12091.htm	09.5	13.4	18.5	22.8	27.3

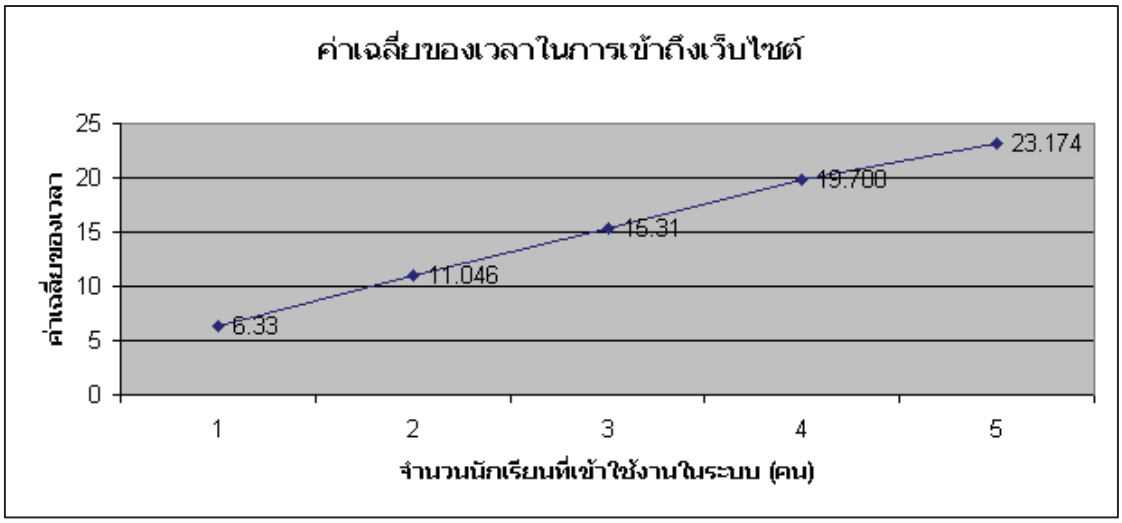
ตารางที่ 25 (ต่อ)

เวลาที่ใช้ในการเข้าถึงข้อมูลเว็บไซต์	จำนวนนักเรียนที่เข้าใช้งานในระบบ				
	10	**30**	**60**	**100**	**150**
http://ss.comparenotebook.info/goURL/180.html	08.5	12.9	17.4	21.7	25.3
http://www8.mobileacce.info/last/13152.html	07.5	11.5	16.8	20.6	24.7
http://avzone.wordpress.com/2009/01/	10.8	14.7	19.1	24.9	28.6
http://www.thaizexstory.com/home/story/128	09.7	13.8	18.8	23.3	27.8
http://thairecent.com/First/2009/392766/	06.5	11.3	15.4	19.2	22.2
http://board.postjung.com/375909.htm	05.4	10.2	14.3	19.2	21.1
http://www.yorkza.com/content/7878/	06.9	11.5	15.8	19.7	22.8
http://www.ryt9.com/s/bmnd/678294	08.3	12.5	17.2	21.5	25.3
http://webboard.yenta4.com/topic/365718	07.5	11.5	15.5	19.8	24.5
รวม	**316.5**	**552.3**	**765.5**	**989.4**	**1158.7**
ค่าเฉลี่ย	**6.33**	**11.046**	**15.31**	**19.788**	**23.174**
ค่าส่วนเบี่ยงเบนมาตรฐาน	**1.8209**	**1.4119**	**1.7698**	**1.6839**	**2.5292**

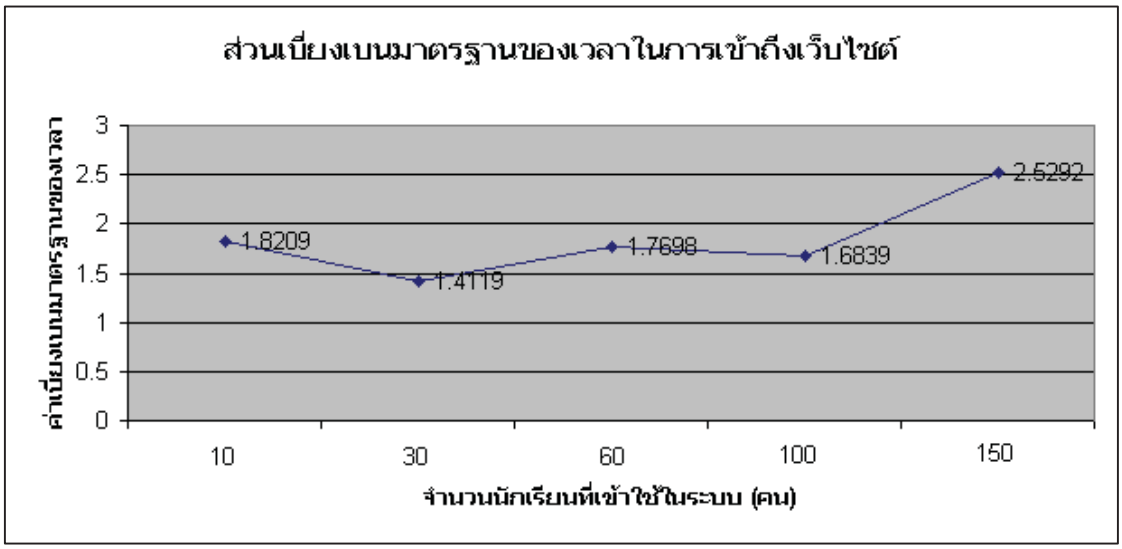


ภาพที่ 36 กราฟแสดงประสิทธิภาพความเร็วของระบบ

จากกราฟจะได้ว่า ปริมาณผู้ใช้งานอินเตอร์เน็ตที่เพิ่มขึ้นเวลาที่ถูกใช้ในการ Block เว็บไซต์จะมีค่ามากขึ้นตามไปด้วย หมายความว่าประสิทธิภาพด้านความเร็วของระบบจะลดลง



ภาพที่ 37 กราฟแสดงค่าเฉลี่ยของเวลาในการเข้าถึงเว็บไซต์



ภาพที่ 38 กราฟแสดงส่วนเบี่ยงเบนมาตรฐานของเวลาในการเข้าถึงเว็บไซต์

เมื่อทำการหาค่าเฉลี่ยพบว่า ปริมาณผู้ใช้งานอินเตอร์เน็ตที่เพิ่มขึ้นค่าเฉลี่ยของเวลาที่ใช้ในการ Block เว็บไซต์จะมีค่ามากขึ้นตามไปด้วย แสดงว่าปริมาณผู้ใช้งานอินเทอร์เน็ตมีผลต่อ

เวลาที่ใช้ในการ Block เว็บไซต์ และเมื่อทำการหาค่าเบี่ยงเบนมาตรฐานพบว่าการกระจายตัวของข้อมูลในแต่ละช่วงของปริมาณผู้ใช้งานอินเตอร์เน็ต 10-100 คน ไม่ต่างกันมากนั้น ซึ่งระบบสามารถรองรับการทำงานได้ดี แต่เมื่อผู้ใช้งานมากขึ้นในระดับ 150 คนขึ้นไป ปริมาณผู้ใช้งานอินเตอร์เน็ตเริ่มมีผลต่อเวลาในการ Block เว็บไซต์มากขึ้น อาจทำให้ระบบการคัดกรองเว็บไซต์ทำงานได้ช้าลง

7.3 ประสิทธิภาพในการรองรับเว็บ 2.0

ในปัจจุบันนี้WEB 2.0 เป็นเว็บไซต์ที่นักพัฒนาให้ความสนใจ และผู้ใช้ก็นิยมใช้บริการเว็บไซต์ที่มีลักษณะเป็นWEB 2.0 มากขึ้น

ลักษณะ WEB 2.0 จะเน้นการแชร์ข้อมูล ความรู้ ความบันเทิงแก่กัน ทำให้เกิดปรากฏการณ์การเกิดชุมชนออนไลน์ ซึ่งกลายเป็นรูปแบบของสังคมประเภทหนึ่งที่อยู่ในโลกอินเทอร์เน็ต (Social network) จะเน้นให้ความสำคัญกับผู้เข้าชมเว็บไซต์ โดยผู้ที่เข้าใช้งานนั้นจะมีส่วนร่วมกับเว็บนั้น ๆ มากขึ้น และไม่ใช่แค่เพียงแวะเข้ามาเยี่ยมชม หรืออ่านอย่างเดียว แต่ยังมีส่วนร่วมในการสร้างสรรค์ (Co-Creation) ให้กับเว็บไซต์แห่งนั้นอีกด้วย โดยให้ผู้ใช้สามารถโต้ตอบ แสดงความคิดเห็นแบ่งปันข้อมูล เป็นผู้สร้างเนื้อหาและนำเสนอข้อมูลได้ด้วยตนเอง และ Web 2.0 สามารถสร้าง user interface ที่สามารถใช้งานได้ง่ายยิ่งขึ้นและรวดเร็วยิ่งขึ้น

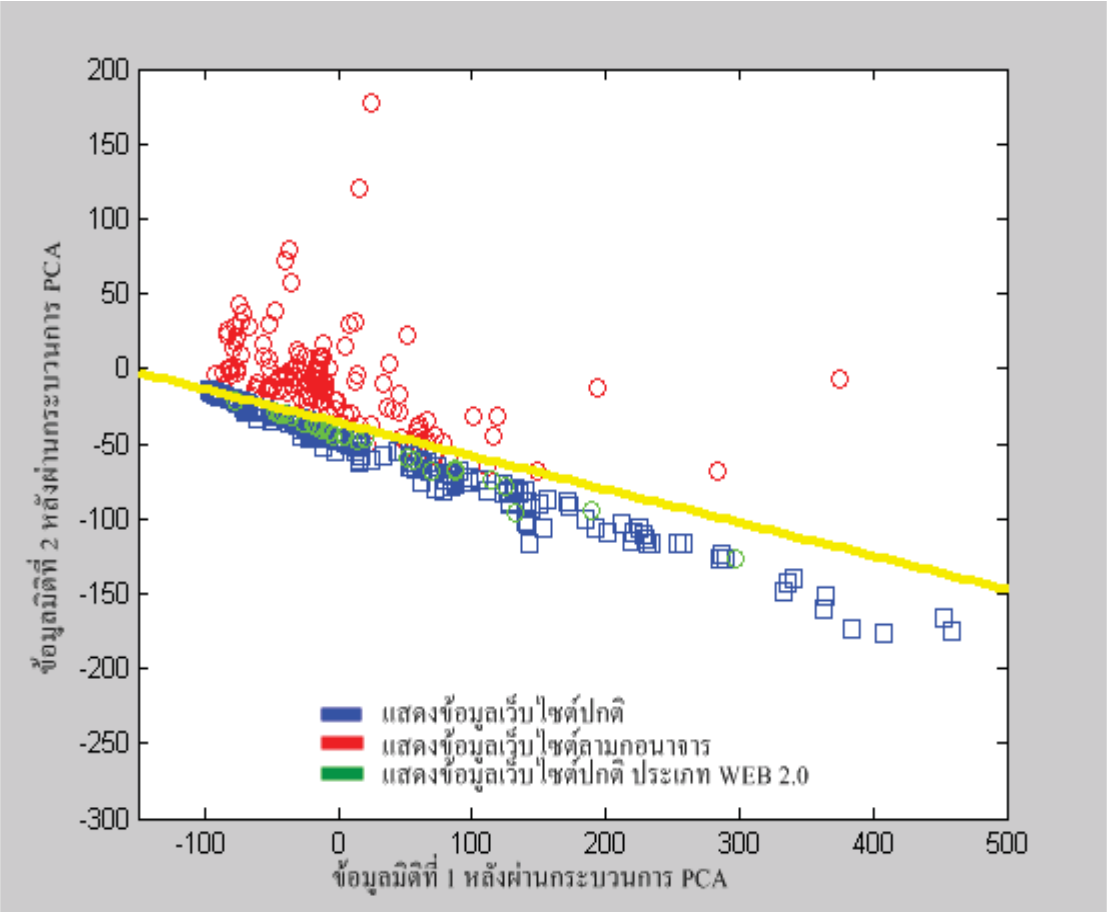
วิทยานิพนธ์ฉบับนี้ได้รองรับเว็บไซต์ประเภท 2.0 หลายประเภท ได้แก่ WORDPRESS, TWITTER, WIKI, BLOGSPOT, EXTEEN และ FLICKR ซึ่งระบบคัดกรองเว็บไซต์สามารถควบคุมการทำงานของเว็บไซต์ 2.0 ได้

ทำการทดลองแบ่งเว็บไซต์ออกเป็นกลุ่มตามแต่ละประเภท โดยทำการคำนวณมิติต่าง ๆ เพื่อใช้ในการทดสอบกับโปรแกรมโดยมีกลุ่มข้อมูลการทดลองดังนี้

	A	B	C	D	E	F	G	H	I	J
1	ประเภทเว็บไซต์	รายชื่อเว็บไซต์	มิติความแตกต่างของเว็บไซต์							
2			จำนวน META	จำนวน A HREF	จำนวน IMG	จำนวน SCRIPT	จำนวน POOL	จำนวนคำลามกอนาจารใน BODY	จำนวน POOL	จำนวนคำลามกอนาจารใน BODY
3	TWITTER	http://twitter.com/traffy	10	115	40	13	0	0	0	0
4		http://twitter.com/okhealthy	10	56	7	13	0	0	0	0
5		http://twitter.com/Prachya	10	86	40	13	0	0	0	0
6		http://twitter.com/JJetrin	10	94	40	13	0	0	0	0
7		http://twitter.com/ggcenter	10	92	40	13	0	0	0	0
8	WORDPRESS	http://yutphuket.wordpress.com/	2	234	45	10	0	0	0	0
9		http://smii.wordpress.com/	6	195	28	14	0	0	0	0
10		http://prachatai.wordpress.com/	2	125	3	8	0	0	0	0
11		http://nham.wordpress.com/2009/01/09/a-new-movie-of-pae-slur/	2	84	15	8	0	0	0	0
12		http://nham.wordpress.com/subscription/	2	64	14	8	0	0	0	0
13	WIKI	http://wiki.it.kmitl.ac.th/Main_Pag	5	61	2	7	0	0	0	0
14		http://wiki.opentle.org/TLE-Live	4	53	1	8	0	0	0	0
15		http://wiki.nectec.or.th/nectecpedia/index.php/Main_Page	3	58	1	6	0	0	0	0
16		http://wiki.tirkx.com/index.php/Main_Page	3	306	17	7	0	0	0	0
17		http://wiki.it.kmitl.ac.th/Introduction_Ubuntu_Software	5	53	2	8	0	0	0	0
18	BLOGSPOT	http://wowboom.blogspot.com/	6	219	280	28	0	0	0	0
19		http://giftshopsale.blogspot.com/	10	19	14	9	0	0	0	0
20		http://d60clubcaker.blogspot.com/	12	194	39	22	0	0	0	0
21		http://fastmp.blogspot.com/	4	83	10	16	0	0	0	0
22		http://klonsiam.blogspot.com/	4	106	25	27	0	0	0	0
23	EXTEEN	http://bignose.exteen.com/	2	126	3	3	0	0	0	0
24		http://darkygirl.exteen.com/	1	226	22	6	0	0	0	0
25		http://phuphu.exteen.com/	1	419	56	17	0	0	0	0
26		http://iphone.exteen.com/	1	91	8	8	0	0	0	0
27		http://thetong.exteen.com/	1	155	54	4	0	0	0	0
28	FLICKR	http://www.flickr.com/photos/thaigow/	5	105	38	19	0	0	0	0
29		http://www.flickr.com/photos/doglookplane/sets	5	71	40	13	0	0	0	0
30		http://www.flickr.com/photos/suvithomesale	5	94	60	16	0	0	0	0
31		http://www.flickr.com/photos/arthit	5	158	69	20	0	0	0	0
32		http://www.flickr.com/photos/sekoser/4315380364/	7	168	115	59	0	0	0	0

ภาพที่ 39 กลุ่มตัวอย่าง Web 2.0 เว็บไซต์ปกติ สำหรับการทดสอบเว็บไซต์ลามกอนาจาร

ผลการทดลองสามารถคัดกรอง WEB 2.0 ได้ โดยเว็บไซต์ที่นำมาทำการทดลองนั้นเป็นเว็บไซต์ปกติและผลการทดสอบที่ได้คือ ตำแหน่ง WEB 2.0 อยู่ใต้เส้นกราฟ ซึ่งมีค่า $C \leq -35.9028$ สรุปได้ว่าเป็นเว็บไซต์ปกติ

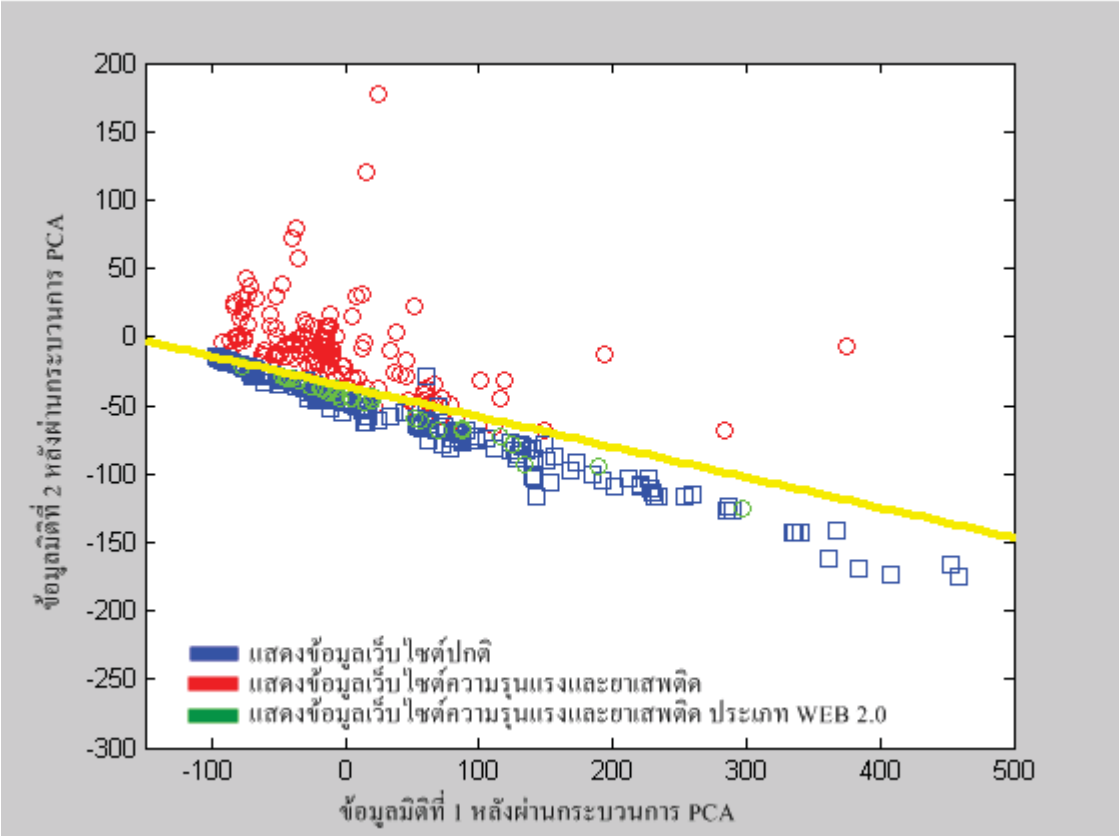


ภาพที่ 40 แสดงถึงระบบสามารถตรวจสอบ WEB 2.0 ประเภทลามกอนาจารได้

	A	B	C	D	E	F	G	H	I	J
1	ประเภทเว็บไซต์	รายชื่อเว็บไซต์	มิติความแตกต่างของเว็บไซต์							
2			จำนวน META	จำนวน A HREF	จำนวน IMG	จำนวน SCRIPT	จำนวน POOL	จำนวนคำสามกอนาจารใน BODY	จำนวน POOL	จำนวนคำสามกอนาจารใน BODY
3	TWITTER	http://twitter.com/traffy	10	115	40	13	0	0	0	0
4		http://twitter.com/okhealthy	10	56	7	13	0	0	0	0
5		http://twitter.com/Prachya	10	86	40	13	0	0	0	0
6		http://twitter.com/JJetrin	10	94	40	13	0	0	0	0
7		http://twitter.com/ggcenter	10	92	40	13	0	3	0	3
8	WORDPRESS	http://yutphuket.wordpress.com/	2	234	45	10	0	0	0	0
9		http://smii.wordpress.com/	6	195	28	14	0	0	0	0
10		http://prachatai.wordpress.com/	2	125	3	8	0	2	0	2
11		http://nham.wordpress.com/2009/01/09/a-new-movie-of-pae-slur/	2	84	15	8	0	0	0	0
12		http://nham.wordpress.com/subscription/	2	64	14	8	0	0	0	0
13	WIKI	http://wiki.it.kmitl.ac.th/Main_Pag	5	61	2	7	0	0	0	0
14		http://wiki.opentle.org/TLE-Live	4	53	1	8	0	0	0	0
15		http://wiki.nectec.or.th/nectecpedia/index.php/Main_Page	3	58	1	6	0	0	0	0
16		http://wiki.tirkx.com/index.php/Main_Page	3	306	17	7	0	0	0	0
17		http://wiki.it.kmitl.ac.th/Introduction_Ubuntu_Software	5	53	2	8	0	0	0	0
18	BLOGSPOT	http://wowboom.blogspot.com/	6	219	280	28	0	2	0	2
19		http://giftshopsale.blogspot.com/	10	19	14	9	0	0	0	0
20		http://d60clubcaker.blogspot.com/	12	194	39	22	0	0	0	0
21		http://fastmp.blogspot.com/	4	83	10	16	0	0	0	0
22		http://klonsiam.blogspot.com/	4	106	25	27	0	0	0	0
23	EXTEEN	http://bignose.exteen.com/	2	126	3	3	0	0	0	0
24		http://darkygirl.exteen.com/	1	226	22	6	0	1	0	1
25		http://phuphu.exteen.com/	1	419	56	17	0	1	0	1
26		http://iphone.exteen.com/	1	91	8	8	0	0	0	0
27		http://thetong.exteen.com/	1	155	54	4	0	0	0	0
28	FLICKR	http://www.flickr.com/photos/thaigow/	5	105	38	19	0	0	0	0
29		http://www.flickr.com/photos/doglookplane/sets	5	71	40	13	0	0	0	0
30		http://www.flickr.com/photos/suvithomesale	5	94	60	16	0	0	0	0
31		http://www.flickr.com/photos/arthit	5	158	69	20	0	0	0	0
32		http://www.flickr.com/photos/sekoser/4315380364/	7	168	115	59	0	0	0	0

ภาพที่ 41 กลุ่มตัวอย่าง Web 2.0 เว็บไซต์ปกติสำหรับการทดสอบเว็บไซต์ความรุนแรง ยาเสพติด

ผลการทดลองสามารถคัดกรอง WEB 2.0 ได้ โดยเว็บไซต์ที่นำมาทำการทดลองนั้นเป็นเว็บไซต์ปกติและผลการทดสอบที่ได้คือ ตำแหน่ง WEB 2.0 อยู่ใต้เส้นกราฟ ซึ่งมีค่า $C \leq -35.9201$ สรุปได้ว่าเป็นเว็บไซต์ปกติ

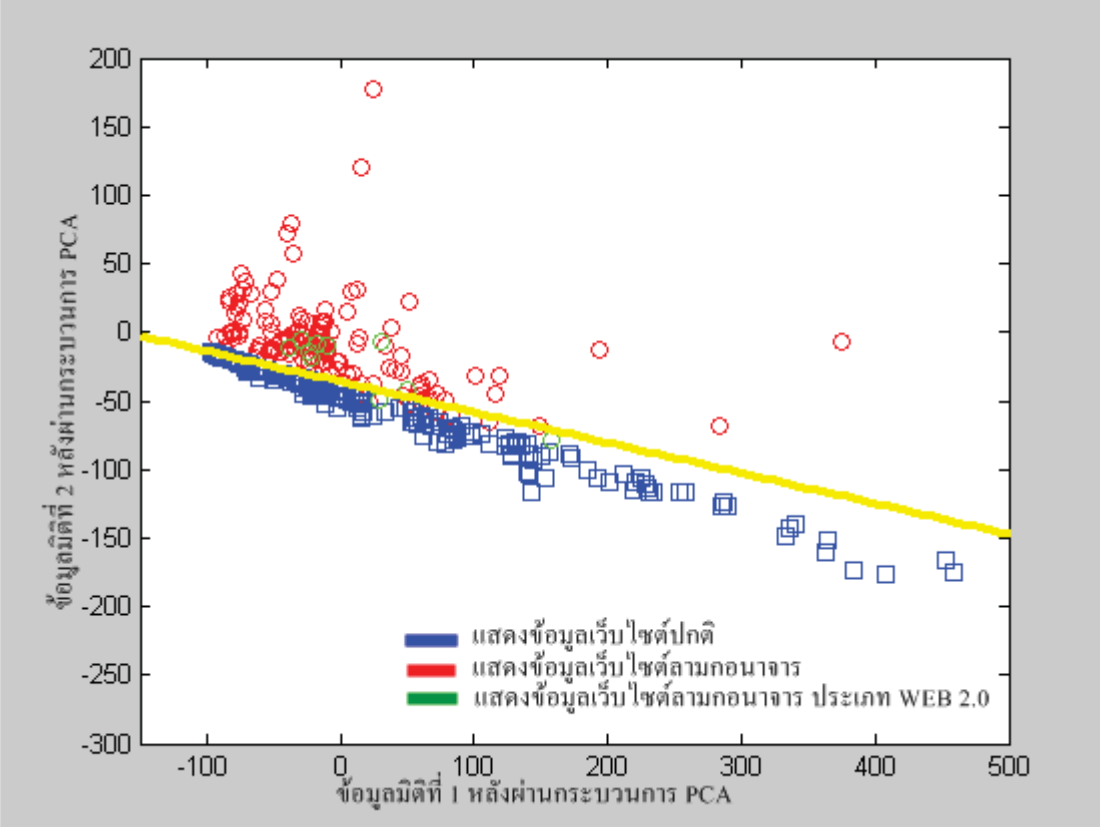


ภาพที่ 42 แสดงถึงระบบสามารถตรวจสอบ WEB 2.0 ประเภทความรุนแรง ยาเสพติดได้

	A	B	C	D	E	F	G	H	I
1		มิติความแตกต่างของเว็บไซต์							
2	รายชื่อเว็บไซต์	จำนวน META	จำนวน A HREF	จำนวน IMG	จำนวน SCRIPT	จำนวน POOL	จำนวนคำสามกอนาจารใน BODY	จำนวน POOL	จำนวนคำสามกอนาจารใน BODY
3	http://twitter.com/yedke	10	76	11	13	0	20	0	20
4	http://ciel.exteen.com/20060101/entry	1	58	59	4	0	22	0	22
5	http://peach69.exteen.com/20071029/censorship-cause-goodness	1	53	38	4	0	15	0	15
6	http://thunska.exteen.com/20060728/entry	1	147	63	4	0	11	0	11
7	http://watchi.exteen.com/20080204/av	1	259	123	4	0	9	0	9
8	http://gorrilaz.exteen.com/200606	1	116	167	5	0	7	0	7
9	http://hostfantasy.exteen.com/20090201/entry	1	69	62	2	0	14	0	14
10	http://69story.wordpress.com/	2	85	7	23	0	20	0	20
11	http://avzone.wordpress.com/	2	124	12	8	0	30	0	30
12	http://yedke.wordpress.com/	2	72	4	8	0	14	0	14

ภาพที่ 43 กลุ่มตัวอย่าง Web 2.0 เว็บไซต์ลามกอนาจารสำหรับการทดสอบเว็บไซต์ลามกอนาจาร

ผลการทดลองสามารถคัดกรอง WEB 2.0 ได้ โดยเว็บไซต์ที่นำมาทำการทดลองนั้นเป็นเว็บไซต์ลามกอนาจารและผลการทดสอบที่ได้คือ ตำแหน่ง WEB 2.0 อยู่เหนือเส้นกราฟ ซึ่งมีค่า C > -35.9028 สรุปได้ว่าเป็นเว็บไซต์ลามกอนาจาร



ภาพที่ 44 แสดงถึงระบบสามารถตรวจสอบ WEB 2.0 ประเภทลามกอนาจารได้

จากผลการทดลองจะได้ว่าโปรแกรมสามารถหาจำนวนคำได้ตามที่กำหนด แต่มีเว็บไซต์บางประเภทไม่สามารถหาคำได้ คือ FACEBOOK จากการตรวจสอบพบว่า Browser ของ FACEBOOK ไม่สนับสนุนการร้องขอข้อมูลของโปรแกรม

บทที่ 5

สรุป อภิปรายผลและข้อเสนอแนะ

จากการพัฒนาระบบป้องกันการเข้าถึงเว็บไซต์ไม่เหมาะสม ทำให้ได้ระบบที่สามารถใช้งานได้จริงและสามารถประยุกต์ใช้ได้ในอนาคต โดยสามารถกำหนดค่าสภาวะแวดล้อมของระบบได้เอง เช่น กำหนดกลุ่มคำพิเศษที่ใช้จำแนก กำหนดเกณฑ์ของการจำแนก ทั้งนี้กล่าวโดยสรุปได้ดังนี้

การบรรลุวัตถุประสงค์การวิจัย

การพัฒนาระบบคัดกรองเว็บไซต์ที่ไม่เหมาะสมในมัธยมศึกษาได้กำหนดวัตถุประสงค์ไว้ คือ

1. เพื่อศึกษาวิธีการคัดกรองเนื้อหาของเว็บไซต์ที่ไม่เหมาะสม

งานวิจัยนี้ได้ทำการศึกษาวิธีการแบ่งกลุ่มข้อมูลเพื่อใช้ในการพัฒนาระบบการคัดกรองเว็บไซต์ที่ไม่เหมาะสม โดยเลือกใช้วิธี PCA (PCA : Principal Component Analysis) ซึ่งเป็นวิธีการเลือกเฉพาะองค์ประกอบที่สำคัญและสามารถตัดองค์ประกอบที่ไม่จำเป็นออกไปได้ ทำให้เหลือแต่องค์ประกอบที่มีความสำคัญในการคัดกรองเว็บไซต์ เพื่อนำไปผ่านกระบวนการแบ่งข้อมูลต่อไป โดยทำการศึกษาวิธี PCA และทำการเปรียบเทียบกับวิธี SVM (SVM : Support Vector Machine) ซึ่งวิธี SVM นั้นเป็นวิธีการแบ่งกลุ่มข้อมูลที่มีประสิทธิภาพที่ได้รับการยอมรับโดยทั่วไป ผลการทดสอบกับจำนวนเว็บไซต์ 400 เว็บไซต์ ปรากฏว่าการใช้ทฤษฎีการวิเคราะห์องค์ประกอบหลักมีค่าความถูกต้องคิดเป็นร้อยละ 92 มากกว่าทฤษฎีซัพพอร์ตเวกเตอร์แมชชีน ซึ่งมีค่าความถูกต้องคิดเป็นร้อยละ 90

2. เพื่อพัฒนาระบบคัดกรองเนื้อหาของเว็บไซต์

หลังจากที่ศึกษาทฤษฎี PCA (PCA : Principal Component Analysis) แล้ว ทำให้สามารถตัดองค์ประกอบที่ไม่จำเป็นเหลือแต่องค์ประกอบที่สำคัญที่ใช้ในการพิจารณาการแบ่งกลุ่มข้อมูลของเว็บไซต์ที่ไม่เหมาะสมออกจากเว็บไซต์ปกติได้แล้ว

จึงนำทฤษฎี PCA มาประยุกต์ใช้ในการสร้างระบบคัดกรองเว็บไซต์ที่ไม่เหมาะสม โดยทำการสร้าง Model รูปแบบข้อมูลการเรียนรู้ให้กับคอมพิวเตอร์ เพื่อให้คอมพิวเตอร์สามารถคัดกรองเว็บไซต์ได้เมื่อมี User ทำการร้องขอเว็บไซต์ ซึ่งรูปแบบ Model ที่ทำการสร้างนี้แบ่งเป็น

2 ประเภท คือ ประเภทที่ 1 Model การคัดกรองเว็บไซต์ไม่เหมาะสมประเภทกลุ่มคำรุนแรงและยาเสพติด ประเภทที่ 2 คือ การคัดกรองเว็บไซต์ไม่เหมาะสมประเภทคำลามกอนาจาร เมื่อสร้าง Model เรียบร้อยแล้ว ได้ทำการพัฒนาโปรแกรมเพื่อติดตั้งระบบบนเครื่องแม่ข่ายอินเตอร์เน็ต เพื่อนำระบบไปติดตั้งที่โรงเรียนศรีวิชัย และสุดท้ายทำการพัฒนาระบบตรวจสอบแก้ไขข้อมูลซึ่งพัฒนาในลักษณะ GUI (GUI : Graphic User Interface) เพื่อให้ผู้ดูแลระบบสามารถตรวจสอบและแก้ไขข้อมูลได้

3. เพื่อประเมินผลระบบที่พัฒนาขึ้น

หลังจากที่ศึกษาทฤษฎี PCA (PCA : Principal Component Analysis) และทำการพัฒนาระบบการคัดกรองเว็บไซต์เรียบร้อยแล้ว ได้ทำการทดสอบประสิทธิภาพการทำงานของระบบ ทั้ง 3 ด้านดังนี้

3.1 ประสิทธิภาพด้านความถูกต้องของการจำแนกเว็บไซต์จากจำนวนเว็บไซต์ทั้งหมด 431 เว็บไซต์ สามารถจำแนกได้ถูกต้อง จำนวน 365 เว็บไซต์ คิดเป็นร้อยละ 84.69 ของจำนวนเว็บไซต์ทั้งหมด

3.2 ประสิทธิภาพด้านความเร็วโดยปัจจัยที่มีผลกระทบต่อความเร็วของระบบนั้นมีหลายปัจจัย เช่น ความเร็วในการเชื่อมต่ออินเตอร์เน็ต จำนวนกลุ่มคำที่ไม่เหมาะสม การคำนวณความถี่ของจำนวนกลุ่มคำเพิ่มเติม แต่ปัจจัยที่มีผลกระทบต่อระบบมากที่สุดคือ ปริมาณผู้ใช้งานอินเทอร์เน็ต ซึ่งเป็นปัจจัยหนึ่งที่ใช้ในการทดสอบประสิทธิภาพด้านความเร็ว เมื่อผู้ใช้งานอินเทอร์เน็ตเพิ่มขึ้นเวลาที่ถูกใช้จะมีค่ามากขึ้นตามไปด้วย

3.3 ประสิทธิภาพด้านการรองรับเว็บ 2.0 จากการทดสอบพบว่าระบบสามารถการคัดกรองเว็บไซต์ที่ไม่เหมาะสมประเภท เว็บ 2.0 ได้

ปัญหาและอุปสรรค

การพัฒนาระบบเกิดปัญหาและอุปสรรค ดังนี้

1. บางเว็บไซต์ไม่เปิดเผย HTML Code ทำให้ไม่สามารถตรวจสอบคำที่ไม่เหมาะสมที่จะนำมาใช้ในการทดลอง เช่น เว็บไซต์ FACEBOOK ซึ่งเป็นประเภท WEB 2.0

2. ระบบที่สร้างสามารถใช้ได้กับเว็บไซต์ภาษาไทยเท่านั้น

3. บางเว็บไซต์ตอบสนองต่อคำร้องขอข้อมูลช้า ทำให้ระบบทำการจำแนกช้าไปด้วย ทั้งนี้ขึ้นอยู่กับปริมาณการใช้งานระบบอินเทอร์เน็ตในเวลานั้นๆด้วย

4. การคัดกรองเว็บไซต์ที่ไม่เหมาะสมของระบบนั้นปัจจัยเรื่องคำเป็นปัจจัยสำคัญที่ใช้จำแนกเว็บไซต์ โดยระบบจะทำการขอข้อมูล Tag HTML โค๊ด มาตรวจสอบ ซึ่งบางเว็บไซต์เก็บคำต่าง ๆ

ไว้บนฐานข้อมูล ทำให้ระบบไม่สามารถตรวจสอบคำได้ อาจทำให้เกิดความคลาดเคลื่อนในการคำนวณค่าเพื่อใช้คัดกรองเว็บไซต์

5. เว็บไซต์บางเว็บไซต์ระบบสามารถจำแนกเว็บไซต์ไม่เหมาะสมออกจากเว็บไซต์ปกติได้แต่มีปัญหาหลังจากส่ง URL ให้ Squid ทำการ Block เว็บไซต์ โดย Squid ไม่สามารถ Block เว็บไซต์ได้ เพราะเว็บไซต์นั้นมีเครื่องหมาย “?” และ “=” ซึ่งตรงกับเครื่องหมายที่เป็นคำสั่งของ Squid ทำให้ Squid ไม่สามารถ Block เว็บไซต์ได้

ข้อเสนอแนะ

การพัฒนาระบบป้องกันการเข้าถึงเว็บไซต์ไม่เหมาะสม เป็นระบบที่พัฒนาขึ้นสำหรับติดตั้งบนเครื่องแม่ข่ายอินเทอร์เน็ตโรงเรียนศรีวิชัยวิทยา ซึ่งใช้ระบบปฏิบัติการ Linux Ubuntu 9.04 ร่วมกับ Squid ที่ทำหน้าที่เป็น Proxy Cache Server ซึ่งผู้ใช้งานสามารถกำหนดค่าสภาวะแวดล้อมในการทำงานได้เอง ซึ่งระบบนี้สามารถนำไปพัฒนาต่อในอนาคต ได้ดังนี้

1. พัฒนาค้นหากลุ่มคำที่สามารถให้ผลการจำแนกที่ให้ผลได้ถูกต้องสูงขึ้น

2. พัฒนาค้นหากลุ่มคำที่ให้ผลกับเว็บไซต์ในภาษาอื่นๆ เช่น ภาษาญี่ปุ่น ภาษาอาหรับ ภาษาจีน ภาษาเกาหลี ฯลฯ ซึ่งยังมีเว็บไซต์ไม่เหมาะสมอยู่อีกเป็นจำนวนมาก

3. เป็นต้นแบบในการพัฒนาวิธีการจำแนกชนิดอื่นๆ เช่น การจำแนกโดยใช้รูปภาพ การจำแนกโดยใช้วิธีวิเคราะห์ html link หรือการจำแนกโดยใช้อัลกอริทึมอื่น

4. พัฒนาค้นหากลุ่มคำที่สามารถจำแนกเว็บไซต์ประเภทอื่นๆ เช่น กลุ่มเว็บไซต์การพนัน กลุ่มเว็บไซต์ด้านความปลอดภัย ฯลฯ

5. พัฒนาโปรแกรมเพื่อรองรับในรูปแบบภาษา Tis620 , Window-874

6. พัฒนาโดยใช้ระบบปฏิบัติการอื่นๆ เช่น Window Server2008

ข้อสรุปใหม่ที่ได้จากการพัฒนาระบบการคัดกรองเว็บไซต์

1. พบว่าการใช้ META TAG IMG TAG A HREF TAG SCRIPT TAG TITLE TAG BODY TAG เป็นองค์ประกอบหลักในการจำแนกเว็บไซต์ ด้วยทฤษฎี PCA (Principal Component Analysis) นั้น สามารถ Classify ข้อมูลได้ดีกว่าวิธี SVM (Support Vector Machine) โดย

ประเภทเว็บไซต์ความรุนแรง ยาเสพย์ติดนั้น PCA คัดกรองได้ถูกต้องร้อยละ 89.5 และ SVM คัดกรองได้ถูกต้องร้อยละ 89

ประเว็บไซต์ลามกอนาจารนั้น PCA คัดกรองได้ถูกต้องร้อยละ 94.5 และ SVM คัดกรองได้ถูกต้องร้อยละ 90

2. พบว่าทฤษฎี Principal Component Analysis ใช้คัดกรองเว็บไซต์ที่ไม่เหมาะสม ประเภท WEB 2.0 ได้ ซึ่งปัจจุบันเป็นเว็บไซต์ที่นิยมใช้กันเช่น HI5 FACEBOOK WIKI TWISTER

3. จากการวิจัยองค์ประกอบต่าง ๆ ของเว็บไซต์ทำให้รู้ว่าองค์ประกอบที่มีความสำคัญระหว่างเว็บไซต์ปกติและเว็บไซต์ไม่เหมาะสมนั้นคือองค์ประกอบของคำ ซึ่งเป็นองค์ประกอบที่มีลักษณะไปในทางเดียวกัน คือกรณีเป็นเว็บไซต์ปกติจะพบคำไม่เหมาะสมอยู่น้อยหรือไม่พบเลย แต่ถ้ากรณีเว็บไซต์ไม่เหมาะสมจะพบคำไม่เหมาะสมอยู่มาก

วิธีการใหม่ที่ได้จากการพัฒนาระบบการคัดกรองเว็บไซต์

ในการวิจัยนี้ได้ทำการคัดกรองเว็บไซต์ที่ไม่เหมาะสม 2 วิธี คือ PCA (Principal Component Analysis) และ SVM (Support Vector Machine) พบว่าวิธีการ PCA (Principal Component Analysis) สามารถคัดกรองได้ดีกว่าวิธี SVM (Support Vector Machine) ดังนั้นหากต้องการ คัดกรองเว็บไซต์ โดยใช้ META TAG IMG TAG A HREF TAG SCRIPT TAG TITLE TAG BODY TAG เป็นองค์ประกอบหลักในการจำแนกเว็บไซต์ สามารถนำวิธีการคัดกรองเว็บไซต์โดยใช้ PCA(Principal Component Analysis) มาเป็นอัลกอริทึมในการคัดกรองได้ โดยนำข้อมูล Training เข้าระบบเพื่อให้ระบบได้ทำการวิเคราะห์องค์ประกอบที่สำคัญของข้อมูลเพื่อใช้สร้างไอเกนเวกเตอร์สำหรับ Model การเรียนรู้ให้ระบบสามารถคัดกรองเว็บไซต์ได้ และทำการทดสอบโดยนำเว็บไซต์ปกติและเว็บไซต์ไม่เหมาะสมมาทดสอบกับระบบ ผลปรากฏว่าระบบสามารถแบ่งแยกเว็บไซต์ไม่เหมาะสมออกจากเว็บไซต์ปกติได้

บรรณานุกรม

ภาษาไทย

กำธร สุทธิรัตน์. “ระบบป้องกันการเข้าถึงเว็บที่ไม่เหมาะสม : กรณีศึกษาโรงเรียนเทพมงคลรังสี จังหวัดกาญจนบุรี.” สารนิพนธ์ปริญญามหาบัณฑิต สาขาวิทยาการมคอมพิวเตอร์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร, 2546.

จันทิมา พลพินิจ และอุมาภรณ์ สายแสงจันทร์. “ระบบการจัดกลุ่มเอกสารข้อความอัตโนมัติด้วยซัพพอร์ตเวกเตอร์แมชชีน.” รายงานการวิจัย ได้รับการสนับสนุนจากเงินงบประมาณรายได้ ประจำปี พ.ศ. 2548 มหาวิทยาลัยมหาสารคาม, 2548.(อัดสำเนา)

จุฑาทิพย์ โพธิ์ทอง. “ความรุนแรงที่ใช้ข่าวที่ปรากฏในภาษาของข่าวอาชญากรรมในหนังสือพิมพ์รายวัน.” วิทยานิพนธ์ศิลปศาสตรมหาบัณฑิต สาขาวิชาภาษาไทยเพื่อการสื่อสาร มหาวิทยาลัยสงขลานครินทร์, 2547.

ธีรพงศ์ โหมดหิรัญ. “การแก้ไขปัญหาความกำกวมของคำในภาษาไทยโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน.” วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, 2548.

ธีรภัทร มนตรีศาสตร์. Squid Proxy Caching Server [ออนไลน์]. เข้าถึงเมื่อ 8 พฤศจิกายน 2552. เข้าถึงได้จาก http:///micro.se-ed.com/content/mc205/MC205_181.asp

บุญสูงเทคโนโลยี. Company Profile [ออนไลน์]. เข้าถึงเมื่อ 8 พฤศจิกายน 2552. เข้าถึงได้จาก http://www.betech.co.th/index.htlm

พิชานี ทัศนเสถียร และภักดิ์ทูล ใจทอง. “ระบบตรวจหาข้อมูลอันตรายในเวิร์ลไวด์เว็บ.” ปริญญานิพนธ์ปริญญาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, 2549.

วิกิพีเดีย. เรขาคณิตวิเคราะห์ [ออนไลน์]. เข้าถึงเมื่อ 16 เมษายน 2010. เข้าถึงได้จาก http://th.wikipedia.org/wiki/

ศูนย์กลางชุมนุนคณิตศาสตร์. อสมการและเอกลักษณ์ เกี่ยวกับ convexity [ออนไลน์]. เข้าถึงเมื่อ 16 เมษายน 2010. เข้าถึงได้จาก http://www.mathcenter.net/forum/showthread.php?t=1262

อานนท์ นามสนิท. “การจำแนกกลุ่มเพลงไทยโดยใช้ซัพพอร์ตเวคเตอร์แมชชีนส์.” วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาเทคโนโลยีคอมพิวเตอร์ คุรุศาสตร์อุตสาหกรรม สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2549.

ภาษาต่างประเทศ

Cooley, W.W. and P.R. Lohnes. Multivariate Data Analysis. New York : John Wiley & Sons Inc., 1971.

Family Online Safety Institute. About ICRA [Online]. Accessed 13 February 2008. Available from http://www.fosi.org/icra

FORTINET. Fortiguard Web Filtering [Online]. Accessed 13 February 2008. Available from http://www.fortinet.com/solutions/wcf.html

Ghani, Rayid, Sean Slattery, and Yiming Yang. Hypertest Categorization using Hyperlink Patterns and Meta Data [Online]. Accessed 29 December 2008. Available from http://www.cs.cmu.edu/~rayid/mypapers/hypertest-icm101.ps

iMimic Networking. DataReactor [Online]. Accessed 13 February 2008. Available from http://www.imimic.com/index.html

Kerlinger, F.N. Foundation of Behavioral Research. United States of America : Hort Rinehart and Winson Inc., 1986.

Kim, J.O., and C.W. Mueller. Factor Analysis : Statistical Methods and Practical Issues. Beverley Hills : Sage Publication, 1978.

Lee, Pui Y., Siu C. Hui, and Alvis Cheuk M. Fong. Neural networks for web content Filtering [Online]. Accessed 13 February 2008. Available from http://ieeexplore. ieee.org/ie15/9670/22293/01039832.pdf?arnumber=1039832

Lewis, David D. and Marc Ringuette. A comparison of two learning algorithms for text Categorization [Online]. Accessed 29 December 2008. Available from http:// www.cs.cmu.edu/afs/user/mnr/www/papers/categ.ps

Lindsay, I. Smith. Background Mathematics : A tutorial on Principal Component [Online]. Accessed 11 January 2009. Available from http://www.cs.otago.ac.nz/cosc453/ student_tutorials/principal_components.pdf

Rastogi, Rajeev, and Kyuseok Shim. PUBLIC : A Decision Tree Classifier that Integrates Building and pruning [Online]. Accessed 29 December 2008. Available from http://www.vldb.org/conf/1998/p404.pdf

Reihaneh, Rongbo Du, Safavi-Naini, and Willy Susilo. “Web Filterring Using Text Classification.” Networks 2003 (ICON2003) 11 (Oct 2003) : 325-330.

SafeSurf. The SafeSurf Internet Rating Standard [Online]. Accessed 13 Febuary 2008. Available from http://www.safesurf.com/ssplan.htm

Secure Computing Corporation. Products at a Glance [Online]. Accessed 13 Febuary 2008. Available from http://www.securecomputing.com/index.cfm?skey=496

Shen , Yirong and Jing Jiang. Improving the Performance of Naïve Bayes for Text Classification [Online]. Accessed 29 December 2008. Available from http://nlp.stanford.edu/courses/cs224n/2003/fp/yirong99/report.pdf

SITA. URL Filtering [Online]. Accessed 13 December 2008. Available from http://www.sita.aero/SITA_URL_Filtering.htm

SquidGuard. About squidGuard [Online]. Accessed 13 February 2008. Available from http://www.squidguard.org/about.html

Stevens, J. Applied Multivariate Statistics for the Social Sciences. 3 rd ed. Mahwah, N.J. : Lawrence Erlbaum Associate Inc., 1996.

Wikipedia. Eigenvalue eigenvector and eigenspace [Online]. Accessed 1 March 2010. Availble from http://en.wikipedia.org/wiki/Eigenvalue,_eigenvector_and_eigenspace

_____, Lagrange_multipliers [Online]. Accessed 16 April 2010. Availble from http://en.wikipedia.org/wiki/Lagrange_multipliers

_____. Support Vector Machine [Online]. Accessed 17 April 2009. Availble from http://en.wikipedia.org/wiki/Support_vector_machine

ภาคผนวก

ภาคผนวก ก

คู่มือการใช้งานของระบบการตรวจสอบควบคุมเว็บไซต์

1. องค์ประกอบในการใช้งานโปรแกรม

โปรแกรมระบบป้องกันการเข้าถึงเว็บไซต์ไม่เหมาะสมใช้โปรแกรม GNU C/C++ ในการพัฒนา ทำงานบนเครื่องแม่ข่ายที่ทำงานด้วยระบบปฏิบัติการลีนุกซ์ Ubuntu8.1 เพื่อให้บริการ Proxy Cache Server ด้วยโปรแกรม Squid

การทำงานของโปรแกรมจะอ่านบันทึกการใช้งานของ Squid จากไฟล์ access.log เพื่อนำเว็บไซต์ไปตรวจจำแนก ถ้าเป็นเว็บไซต์ไม่เหมาะสมจะทำการเพิ่มใน Blacklists โดยอัตโนมัติ

2. วิธีการติดตั้งโปรแกรมระบบป้องกันการเข้าถึงเว็บไม่เหมาะสม

2.1 ทำ Transparency เพื่อไม่ให้เครื่องลูกข่ายเข้าใช้งานอินเทอร์เน็ต โดยไม่ผ่าน Proxy โดยแก้ไขไฟล์ ดังนี้

แก้ไขไฟล์ /etc/rc.local เพิ่มข้อความต่อไปนี้

```
iptables -t nat -F
iptables -t mangle -F
iptables -t filter -F
iptables -X
iptables -A FORWARD -j ACCEPT
iptables -t nat -A POSTROUTING -o eth0 -j MASQUERADE
iptables -t nat -A PREROUTING -i eth1 -p tcp --dport 80 -j REDIRECT --to-port 8080
```

แก้ไขไฟล์ /etc/squid/squid.conf เพิ่มข้อความต่อไปนี้

```
http_port 8080 transparent
```

2.2 สร้างระบบ Blacklists โดยแก้ไขไฟล์ /etc/squid/squid.conf โดยเพิ่มข้อความต่อไปนี้

```
acl  ProxyFilter  src  127.0.0.1/255.255.255.255
http_access allow  ProxyFilter
acl lock1 url_regex '/home/ProxyFilter/blacklist1.txt'
http_access deny lock1
deny_info http://www.math26.com/thesis/warning1.html lock1
acl lock2 url_regex '/home/ProxyFilter/blacklist2.txt'
http_access deny lock2
deny_info http://www.math26.com/thesis/warning2.html lock2
```

2.3 สร้างระบบ Blacklists โดยแก้ไขไฟล์ /etc/squid/squid.conf โดยเพิ่มข้อความต่อไปนี้ การปรับรูปแบบของไฟล์ /var/log/squid/access.log ใหม่ โดยแก้ไขไฟล์ /etc/squid.conf ในบรรทัดที่มีข้อความว่า

cache_access_log /var/log/squid/access.log

แทนที่เป็น

logformat common %{%Y-%m-%d %H:%M:%S}tl %6tr %>a %Ss/%03Hs %<st %rm %ru %un %Sh/%<A %mt

2.4 สร้างไดเรกทอรี่ /home/ProxyFilter/ แล้วคัดลอกโปรแกรมลงในไดเรกทอรี่ home/ProxyFilter โดยใช้คำสั่งต่อไปนี้

#mount /dev/cdrom /mnt

#cp –prv /mnt/ProxyFilter /home/

#chmod –R 777 /home/ProxyFilter/

2.5 การรันโปรแกรม

#.ProxyFilter –r เคลียร์ค่า blacklist1.txt,blacklist2.txt,whitelist.txt

#./ProxyFilter –s http://www.com-th.net สั่งให้โปรแกรมทำงานโดยตรวจสอบเว็บไซต์ที่ระบุเพียงเว็บไซต์เดียว

#./ProxyFilter รันโปรแกรมใน Daemon Mode(ทำงานตลอดเวลา)

หากต้องการให้โปรแกรมทำงานเองอัตโนมัติทุกครั้งที่มีการเปิดเครื่อง ให้แก้ไขไฟล์ /etc/rc.local โดยเพิ่มข้อความดังต่อไปนี้ต่อท้ายข้อความเดิม

cd /home/ProxyFilter/

./ProxyFilter&

2.6 การปรับค่าสภาวะแวดล้อมในการทำงาน สามารถแก้ไขไฟล์ /home/ProxyFilter/proxy-filter.conf ดังนี้

ACCESS_LOG=/var/log/squid/access.log ไฟล์เก็บบันทึกการใช้งาน Proxy

BLACK_LIST_1=/home/ProxyFilter/blacklist1.txt ไฟล์เก็บ Blacklist1.txt

BLACK_LIST_2=/home/ProxyFilter/blacklist2.txt ไฟล์เก็บ Blacklist2.txt

WHITE_LIST=/home/ProxyFilter/whitelist.txt ไฟล์เก็บ Whitelist.txt

DICTIONARY_1=/home/ProxyFilter/dict1 ไฟล์เก็บกลุ่มคำความรุนแรง

DICTIONARY_2=/home/ProxyFilter/dict2ไฟล์เก็บกลุ่มคำลามกอนาจาร

PROXYPORT=8080 port ที่ใช้ติดต่อกับ Proxy

MAX_COUNT_1=4 เกณฑ์ความถี่ของกลุ่มคำความรุนแรง

MAX_COUNT_2=7 เกณฑ์ความถี่ของกลุ่มคำลามกอนาจาร

3. การติดตั้ง GUI (Graphic User Interface)

mount /dev/cdrom /mnt

cp /mnt/webfilter /var/www

sudo chmod -R 777 /var/www/webinterface

sudo chown -R www-data:www-data /var/www/webinterface /home/ProxyFilter/

chmod 777 /usr/sbin/squiq

แก้ไขไฟล์ /etc/sudoers โดยเพิ่มบรรทัดต่อไปนี้ต่อท้ายบรรทัดสุดท้าย

www-data ALL=(ALL) NOPASSWD:ALL

4. การสั่ง อิมพอร์ตฐานข้อมูล Mysql

mysql -uroot -p < /var/www/squid_admin/MYSQL/squid.sql

รหัสผ่านแอดมิน

User admin

Pass 1234

ใส่รหัสผ่าน Mysql โดยแก้ไขไฟล์ต่อไปนี้ให้ตรงกับที่เราติดตั้ง

#vi /var/www/webinterface/include/config.php

$obj_dbconfig->set_username("root");

$obj_dbconfig->set_password("1234");

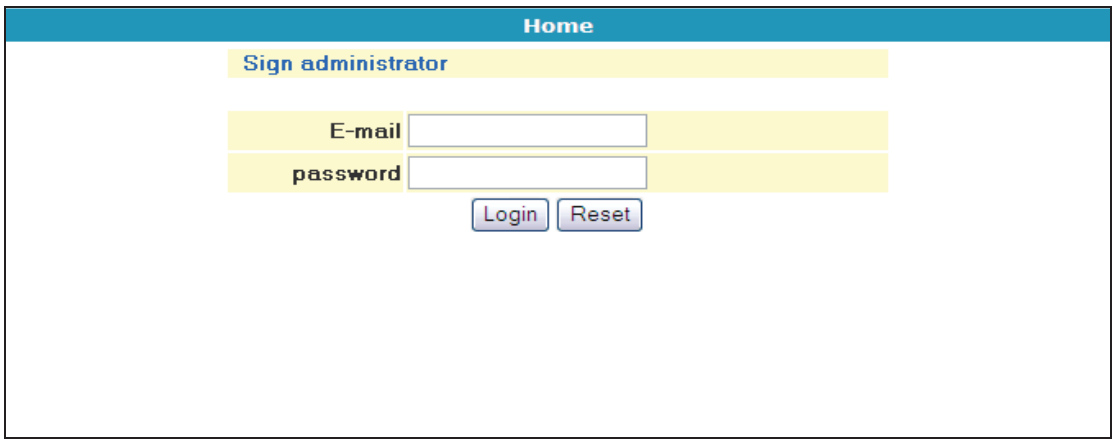
ทดสอบเรียกเว็บไซต์ในที่นี้ให้เครื่องแม่ข่าย มีหมายเลข IP=192.168.1.1

http://l92.168.1.1/webinterface/

5. ขั้นตอนการใช้งานระบบฐานข้อมูลการควบคุมเว็บไซต์

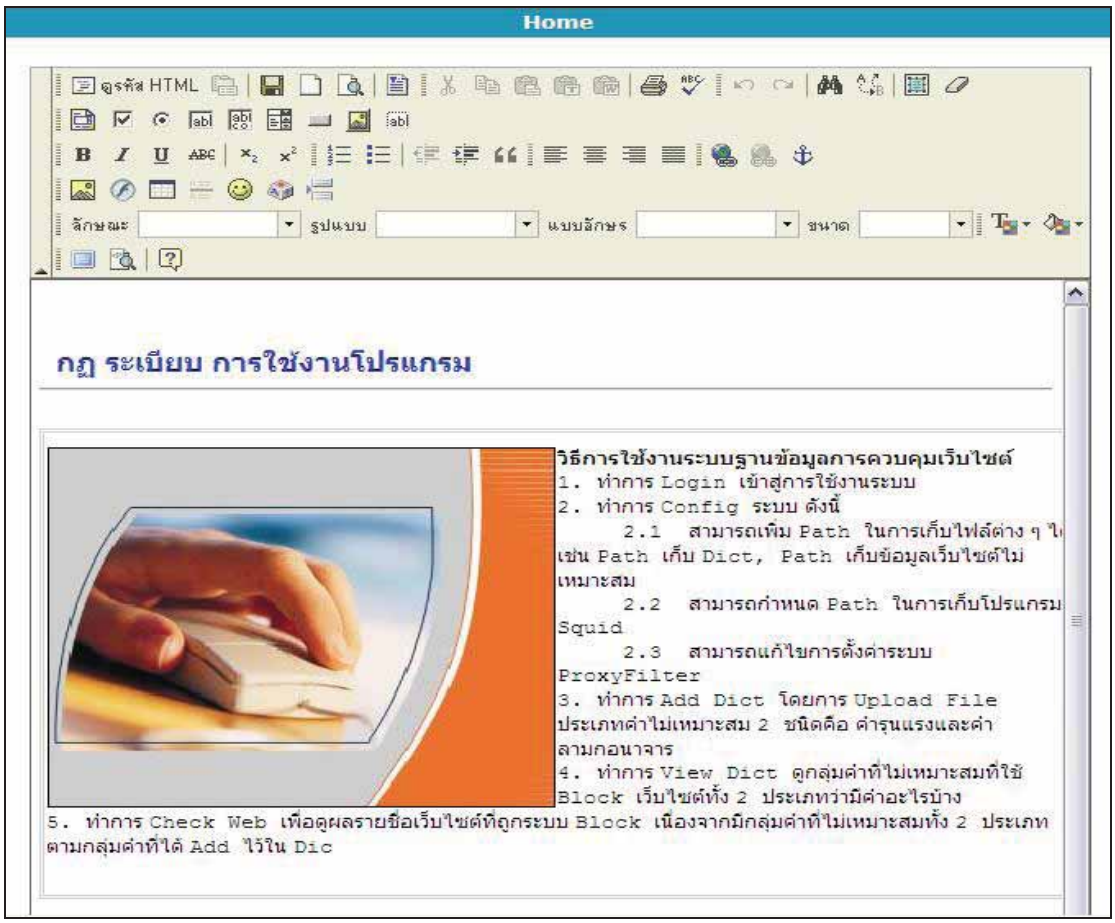
5.1 ผู้ใช้ระบบทำการ Login เข้าสู่ระบบ โดยทำการกรอกข้อมูลดังนี้

Username ที่ช่อง E-mail และ Password ที่ช่อง Password



ภาพที่ 45 หน้าจอ Login

5.2 ผู้ใช้สามารถแจ้งข่าวสารในหน้า Home ผ่าน Menu News



ภาพที่ 46 หน้าจอแจ้งข่าวสารในหน้า Home ผ่าน Menu News

5.3 ผู้ใช้ทำการกำหนดค่าการทำงานต่าง ๆ ผ่านเมนู Config

5.3.1 ทำการเพิ่ม Path File ต่าง ๆ ได้โดยกำหนดผ่านเมนูเพิ่ม Text File ใส่ Path ที่ต้องการเพิ่มเข้าไปในช่องที่อยู่ แล้วกดปุ่มบันทึก

5.3.2 สามารถกำหนด Path File สำหรับการเก็บค่า Config ของโปรแกรม Squid ได้โดยกำหนดผ่านเมนูตั้งค่าระบบ Squid ใส่ Path ที่ต้องการเข้าไปในช่องที่อยู่ แล้วกดปุ่มบันทึก

5.3.3 กำหนดค่า Config ของ ProxyFilter โดยกำหนดผ่านเมนูตั้งค่าระบบ ProxyFilter

5.3.4 กำหนด Path ACCESS_LOG เป็น Path สำหรับการเก็บรายชื่อเว็บไซต์ที่ USER ใช้งานทั้งหมด

5.3.5 กำหนด Path BLACK_LIST_1 เป็น Path สำหรับการเก็บรายชื่อเว็บไซต์ไม่เหมาะสมประเภทความรุนแรง ยาเสพติดที่ USER มีการใช้งาน

5.3.6 กำหนด Path BLACK_LIST_2 เป็น Path สำหรับการเก็บรายชื่อเว็บไซต์ไม่เหมาะสมประเภทลามกอนาจารที่ USER มีการใช้งาน

5.3.7 กำหนด Path WHITE_LIST เป็น Path สำหรับการเก็บรายชื่อเว็บไซต์ปกติที่ USER มีการใช้งาน

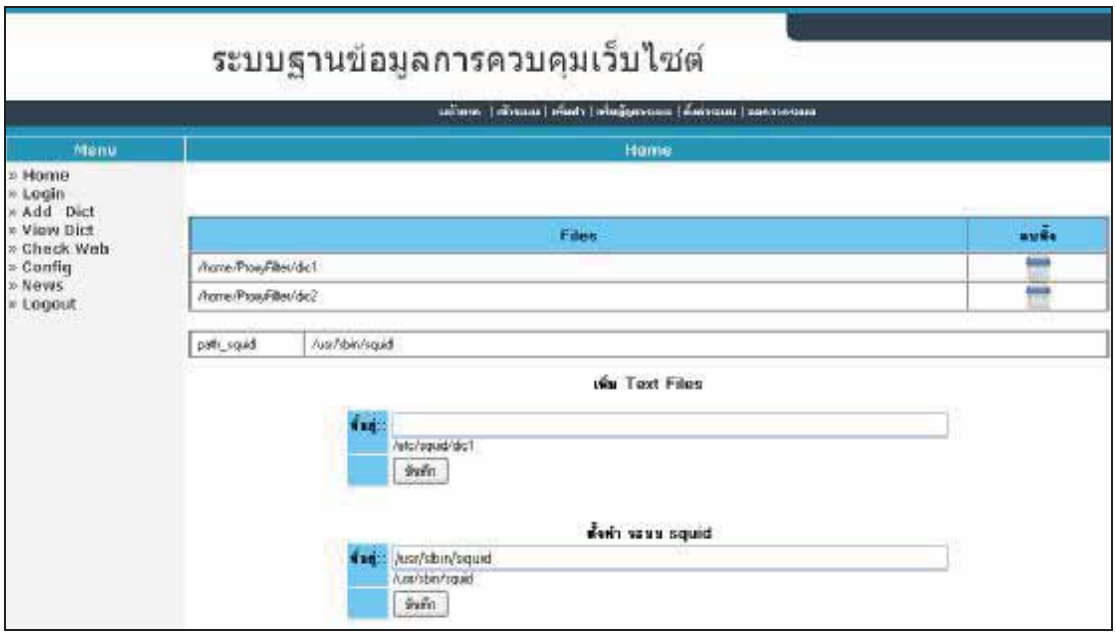
5.3.8 กำหนด Path DICTIONARY_1 สำหรับเก็บกลุ่มคำไม่เหมาะสมที่ใช้ Block เว็บไซต์ประเภทกลุ่มคำรุนแรง ยาเสพติด

5.3.9 กำหนด Path DICTIONARY_2 สำหรับเก็บกลุ่มคำไม่เหมาะสมที่ใช้ Block เว็บไซต์ประเภทกลุ่มคำลามกอนาจาร

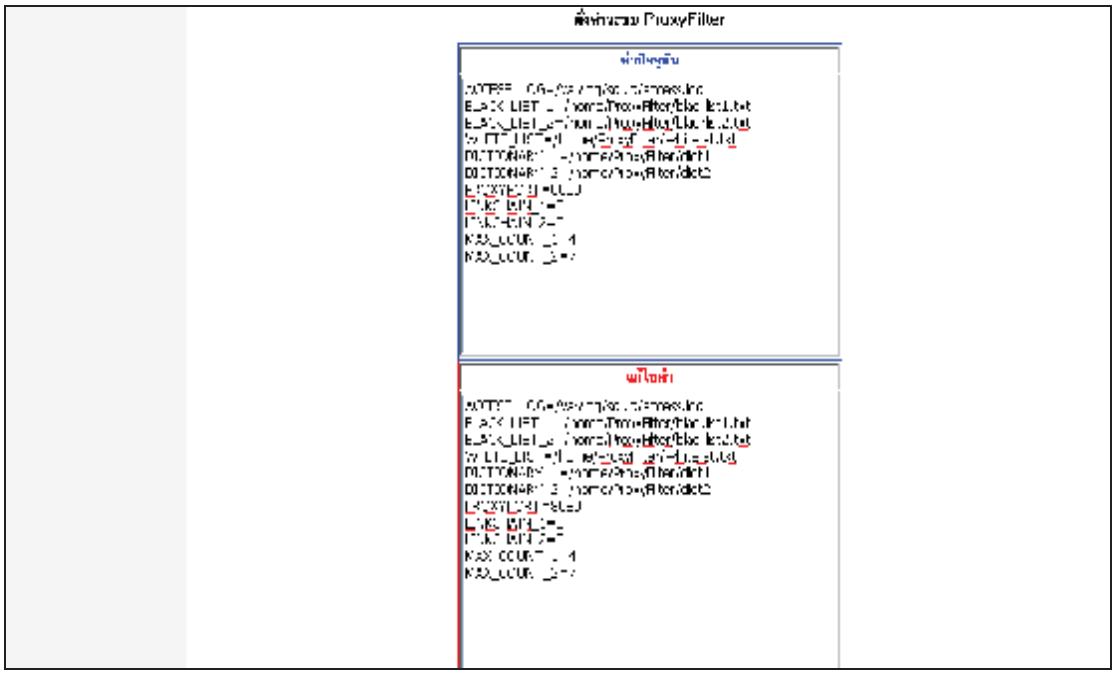
5.3.10 กำหนด Port สำหรับติดต่อกับ Proxy

5.3.11 กำหนด MAX_COUNT_1 จำนวนคำที่เหมาะสมในการ Block เว็บไซต์ความรุนแรง ยาเสพติด

5.3.12 กำหนด MAX_COUNT_2 จำนวนคำที่เหมาะสมในการ Block เว็บไซต์ลามกอนาจาร



ภาพที่ 47 หน้าจอเมนู Config Files

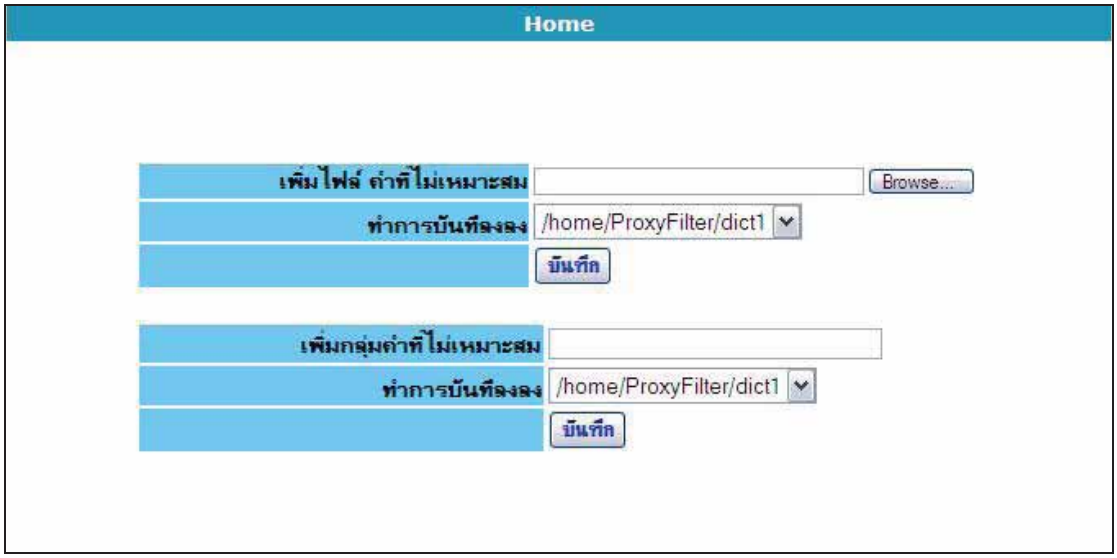


ภาพที่ 48 หน้าจอเมนู Config ProxyFilter

5.4 หน้าจอเพิ่มกลุ่มคำไม่เหมาะสมโดยสามารถเพิ่มกลุ่มคำได้ 2 วิธี

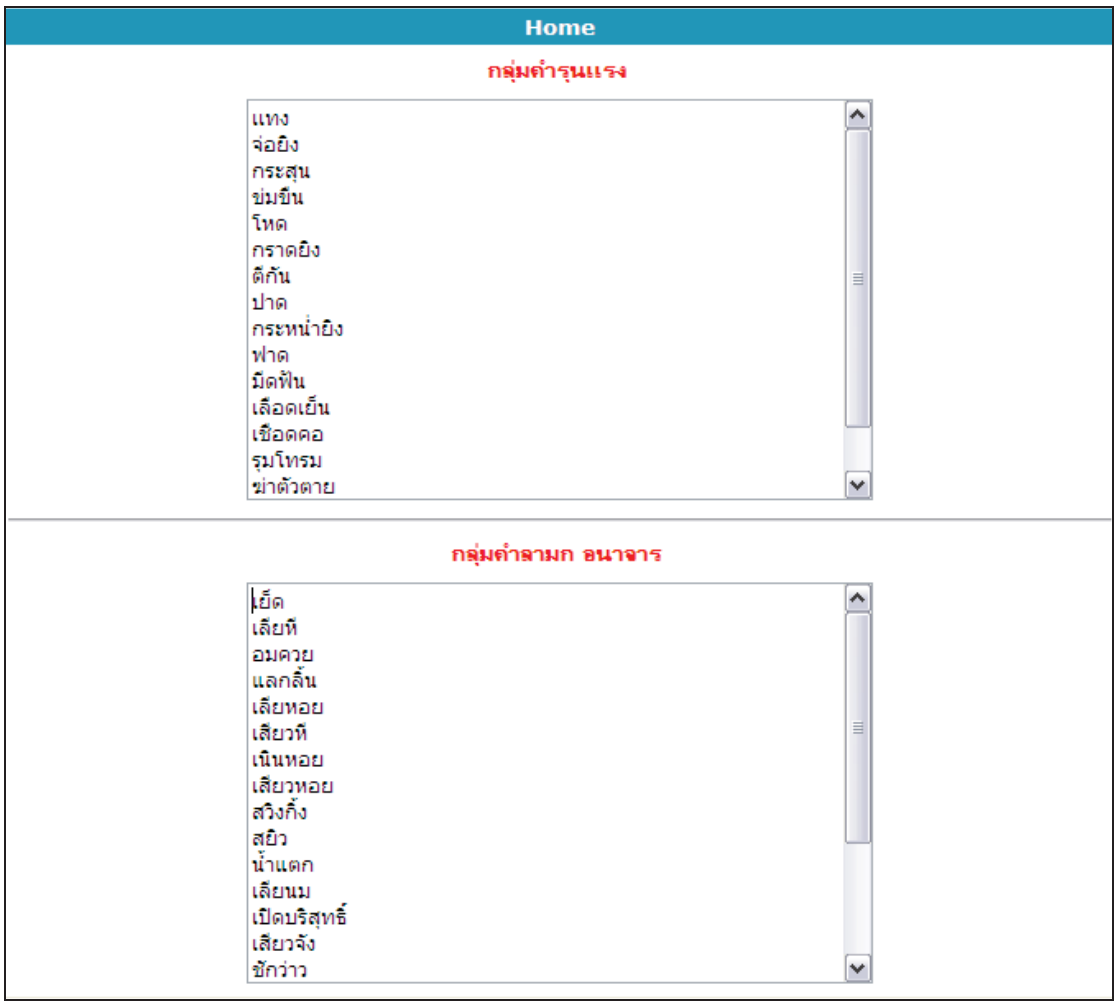
5.4.1 สามารถเพิ่มคำไม่เหมาะสมในรูปแบบ Text File ได้

5.4.2 สามารถเพิ่มกลุ่มคำโดยเพิ่มทีละคำผ่านหน้าจอ



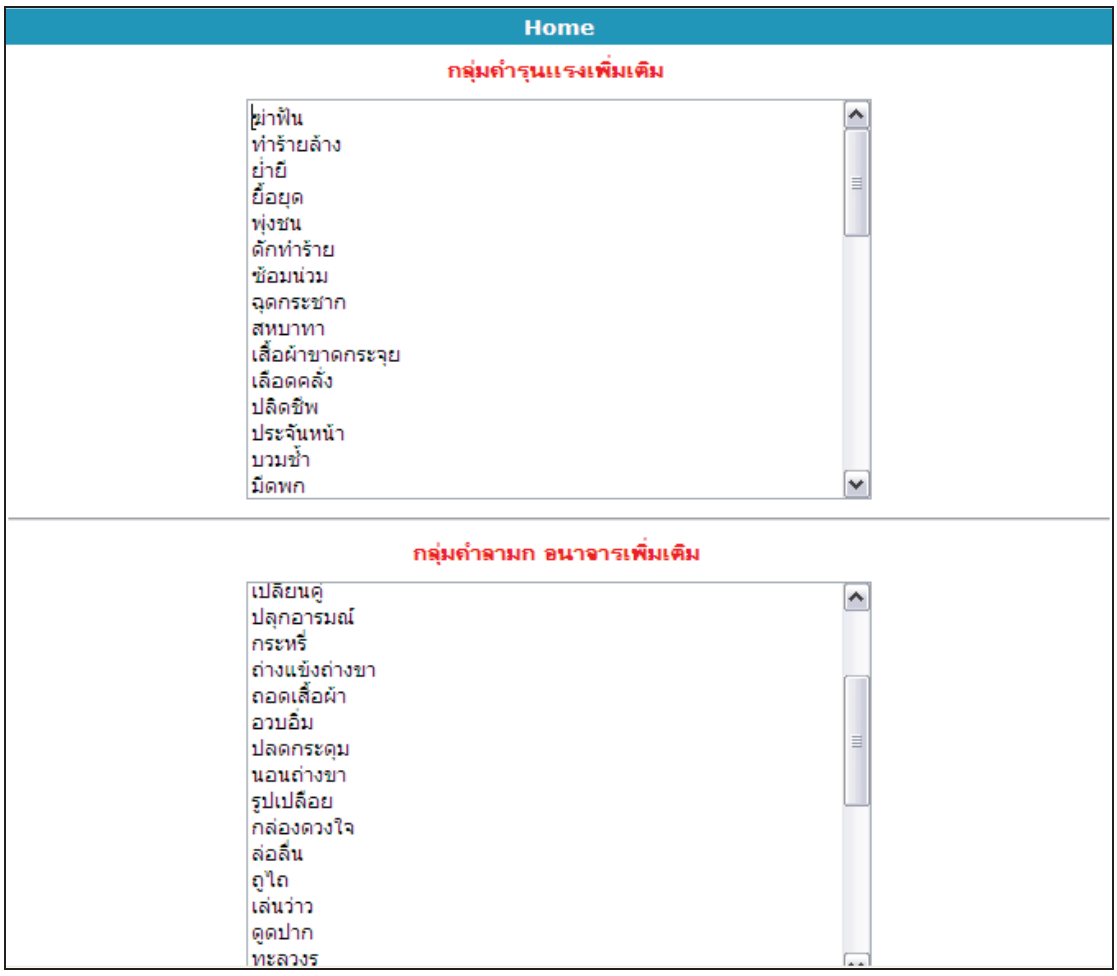
ภาพที่ 49 หน้าจอเพิ่มกลุ่มคำไม่เหมาะสม

5.5 สามารถดูกลุ่มคำที่ไม่เหมาะสมที่ใช้ Block เว็บไซต์ความรุนแรง ยาเสพติดและเว็บไซต์ลามกอนาจารผ่านหน้าจอ View Dict 1-2



ภาพที่ 50 หน้าจอแสดงกลุ่มคำไม่เหมาะสม

5.6 สามารถดูกลุ่มคำที่ไม่เหมาะสมที่อาจเป็นกลุ่มคำที่ใช้ Block เว็บไซต์ผ่านหน้าจอ View Dict 3-4



ภาพที่ 51 หน้าจอแสดงกลุ่มคำไม่เหมาะสมเพิ่มเติม

ภาคผนวก ข

คู่มือการใช้งาน SVM^{light} V6.02

1. เกี่ยวกับ SVM^{light}

SVM^{light} เป็นเครื่องมือตัวหนึ่งที่ใช้ในการคำนวณค่าต่าง ๆ ในอัลกอริทึม SVM ช่วยแก้ปัญหาการแบ่งข้อมูลออกจากกัน สามารถคำนวณประสิทธิภาพการแบ่งข้อมูล ได้แก่ ค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก สามารถคำนวณค่าการแบ่งข้อมูลโดยใช้ทฤษฎี Kernel Functions และเป็นเครื่องมือที่สามารถรองรับการคำนวณข้อมูลได้มาก

SVM^{light} เป็นโปรแกรมที่ติดตั้งง่าย ไม่เสียค่าใช้จ่าย และสามารถใช้งานบน Windows หรือบนระบบปฏิบัติการลีนุกซ์

2. วิธีการติดตั้งโปรแกรม SVM^{light} บน window

เมื่อโหลดโปรแกรม SVM^{light} ลงมาแล้วทำการแตก ZIP ไฟล์ จะได้ไฟล์ .exe 2 ไฟล์ คือ svm_learn และ svm_classify การทำงานของโปรแกรม SVM^{light} จะทำงานผ่าน command line

3. การเตรียมข้อมูลสำหรับการ Training

กำหนดค่าการแบ่งข้อมูลโดยให้คำตอบของค่ากลุ่มที่ 1 เป็น +1 และคำตอบของค่ากลุ่มที่ 2 เป็น -1 โดยกำหนดคำตอบให้เป็นคอลัมน์แรกของข้อมูล และคอลัมน์ต่อไปเป็นมิติความแตกต่างของข้อมูลโดยแสดงลำดับที่ของมิติบวกด้วยเครื่องหมาย ":" และต่อด้วยจำนวนความถี่ของการพบมิตินั้นรูปแบบข้อมูลการนำเข้าระบบ คือ ลำดับของมิติ:ความถี่รวมของค่ามิตินั้น

1	1:5	2:95	3:17	4:12	5:1	6:10	7:1	8:10
1	1:4	2:84	3:3	4:9	5:2	6:27	7:2	8:27
1	1:5	2:36	3:36	4:14	5:1	6:20	7:1	8:20
1	1:1	2:1	3:6	4:0	5:1	6:30	7:1	8:30
1	1:1	2:58	3:1	4:0	5:0	6:10	7:0	8:10
1	1:7	2:145	3:131	4:23	5:1	6:8	7:1	8:8
1	1:3	2:10	3:5	4:23	5:0	6:13	7:0	8:13
1	1:5	2:36	3:34	4:9	5:0	6:20	7:0	8:20
1	1:3	2:17	3:5	4:4	5:2	6:35	7:2	8:35
-1	1:18	2:271	3:150	4:18	5:0	6:6	7:0	8:6
-1	1:3	2:241	3:42	4:14	5:0	6:0	7:0	8:0
-1	1:4	2:472	3:459	4:14	5:0	6:1	7:0	8:1
-1	1:8	2:136	3:217	4:18	5:0	6:10	7:0	8:10
-1	1:3	2:84	3:76	4:2	5:0	6:0	7:0	8:0
-1	1:6	2:302	3:196	4:9	5:0	6:0	7:0	8:0
-1	1:7	2:182	3:169	4:15	5:0	6:0	7:0	8:0
-1	1:1	2:72	3:26	4:11	5:0	6:0	7:0	8:0
-1	1:7	2:62	3:181	4:14	5:0	6:0	7:0	8:0
-1	1:4	2:319	3:214	4:5	5:0	6:0	7:0	8:0
-1	1:13	2:153	3:292	4:20	5:0	6:0	7:0	8:0

ภาพที่ 52 แสดงรูปแบบข้อมูล Data Train

4. คำสั่งการ Training ข้อมูลเพื่อสร้าง Model

คำสั่งในการสร้าง Model มีดังนี้ “svm_learn [option] example_file model_file”

example_file คือ Data Train ที่มีการจัดรูปแบบแล้ว

Model_file คือ โมเดลเรียนรู้ของระบบซึ่งเป็น output ที่ได้

Option ที่ใช้ในการทดสอบมีดังนี้

Learning options:

-z{c,r,p} c หมายถึงการ Classification, r หมายถึงการ regression และ p หมายถึงการ Preference ranking ซึ่งค่าที่เราใช้คือค่า c การ Classification

Kernel Options:

-t int คือ ชนิดของ kernel function

0: linear

1: polynomial

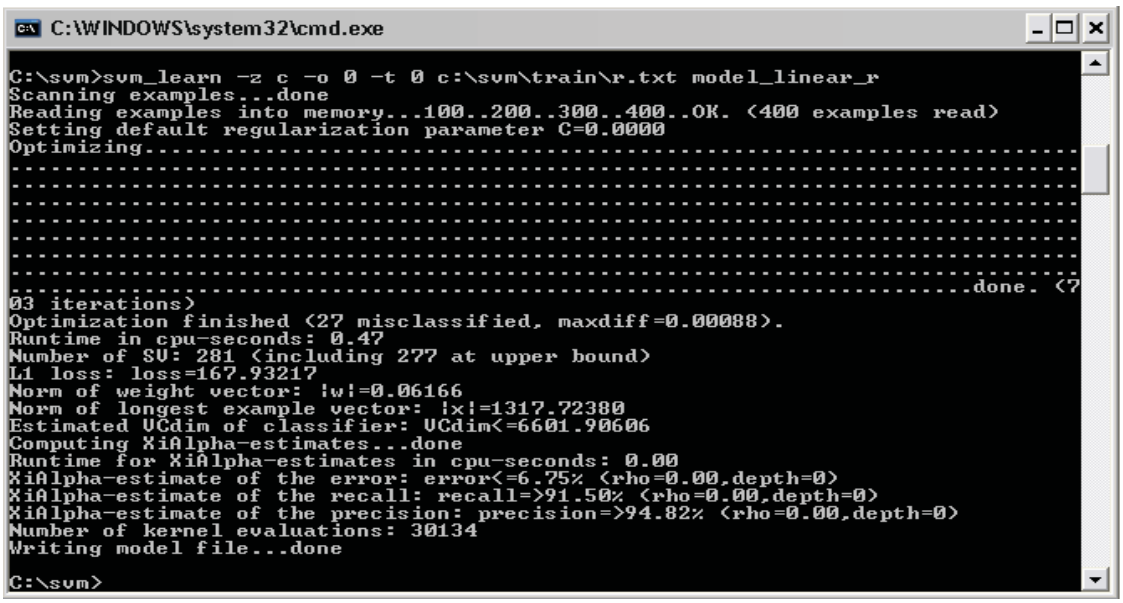
2: radial basis function (rbf)

3: sigmoid

-d int ค่าพารามิเตอร์ d ใน polynomial kernel

-g float ค่าพารามิเตอร์ gamma ใน rbf kernel

-s float ค่าพารามิเตอร์ s ใน sigmoid kernel



ภาพที่ 53 ตัวอย่างการสร้าง model ด้วยวิธี SVM แบบ linear

5. คำสั่งการ Testing ข้อมูลเพื่อทดสอบข้อมูลที่เข้ามาในระบบ

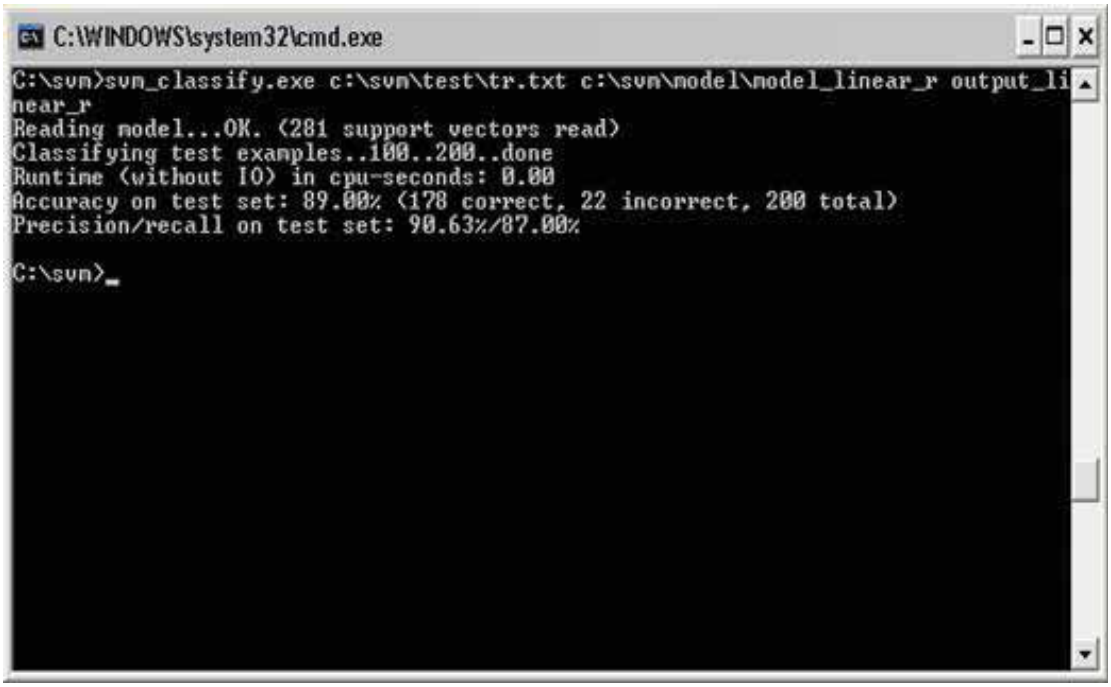
นำข้อมูล Data Test มาทดสอบกับ model การเรียนรู้ของระบบ โดยใช้คำสั่งในการทดสอบระบบดังนี้ “svm_classify example_file model_file output_file” โดย

example_file คือ Data Test ที่มีการจัดรูปแบบแล้ว

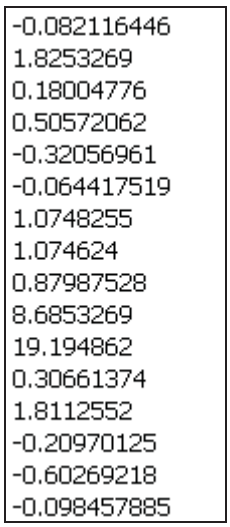
Model_file คือ โมเดลเรียนรู้ของระบบจากการสร้างไว้

Output_file คือ output ข้อมูลที่ได้จากการทดสอบข้อมูล Data Test กับ

ระบบ model การเรียนรู้ ถ้าค่า output ที่ได้แสดงเป็น + แสดงว่าเป็นข้อมูลกลุ่มเดียวกับ +1 ที่กำหนดไว้ตั้งแต่การนำข้อมูลเข้าสร้าง Model ถ้าค่า output ที่แสดงได้เป็น – แสดงว่าเป็นข้อมูลกลุ่มเดียวกับ –1



ภาพที่ 54 ตัวอย่างการทดสอบข้อมูลด้วยวิธี SVM แบบ linear



ภาพที่ 55 ตัวอย่าง output ที่ได้จากการทดสอบ

ภาคผนวก ค

รายชื่อเว็บไซต์ความรุนแรง ยาเสพติด 200 เว็บไซต์

รายชื่อเว็บไซต์ความรุนแรง ยาเสพติด 200 เว็บไซต์ มีดังนี้

1. http://news.mthai.com/general-news/35767.html
2. http://tnews.teenee.com/crime/2351.html
3. http://board.postjung.com/topic-423629.html
4. http://webboard.yenta4.com/topic/318269
5. http://play.kapook.com/vdo/show-26526
6. http://regist53.blogspot.com/2009/08/13.html
7. http://www.spiceday.com/thread-142958-1-2.hot
8. http://atcloud.com/stories/15199
9. http://www2.nurnia.com/5217/08/thai-social-political-economic/
10. http://bbs.soizaa.com/archiver/tid-8228.html
11. http://www.pitakthai.com/social/217.html
12. http://www.ohthai.net/10732.html
13. http://www.ryt9.com/s/bmnd/730195/
14. http://atcloud.com/stories/45137
15. http://www2.nurnia.com/29249/08/thai-social-political-economic/
16. http://www.pitakthai.com/crime/3487.html
17. http://webboard.yenta4.com/topic/251441
18. http://news.mthai.com/general-news/52345.html
19. http://www.ryt9.com/s/bmnd/712912
20. http://www.ryt9.com/s/bmnd/742336
21. http://talk.mthai.com/topic/20383
22. http://atcloud.com/stories/45891
23. http://tnews.teenee.com/crime/27312.html
24. http://www2.nurnia.com/14126/01/thai-social-political-economic/
25. http://talk.mthai.com/topic/44394
26. http://webboard.yenta4.com/topic/48244
27. http://www.click.in.th/news/local/35771.html
28. http://www.huaybondin.net/forum/thread-5301-1-1.html

29. http://www.jobedd.com/th/page/id/744
30. http://www.opensubscriber.com/message/mass-groups@googlegroups.com/11255345.html
31. http://www.yorkza.com/content/8449/
32. http://tnews.teenee.com/crime/583.html
33. http://topsy.com/tb/www.thairath.co.th/content/oversea/51082
34. http://www.opensubscriber.com/message/mass-groups@googlegroups.com/9264396.html
35. http://regist53.blogspot.com/2009/11/25-24-2552.html
36. http://board.siamjung.com/index.php/topic,1249.0.html
37. http://news.ohpicpost.com/%E0%B8%82%E0%B9%88%E0%B8%B2%E0%B8%A7/19651/
38. http://board.postjung.com/423124.html
39. http://thairecent.com/Local/2009/143658/
40. http://www.yorkza.com/content/7979/
41. http://www.huaybondin.net/forum/thread-46840-1-1.html
42. http://www.yorkza.com/content/2605/
43. http://www.zone-it.com/130415
44. http://regist53.blogspot.com/2009/09/3_28.html
45. http://regist53.blogspot.com/2009/04/4_30.html
46. http://www.pochnews.com/article/1451.htm
47. http://www.bangkokhealth.com/index.php/2009-01-19-02-52-08/136-2009-01-19-07-38-29
48. http://thairecent.com/First/2009/392766/
49. http://board.postjung.com/375909.htm
50. http://www.yorkza.com/content/7878/
51. http://www.ryt9.com/s/bmnd/678294
52. http://board.postjung.com/m/375909.html
53. http://webboard.yenta4.com/topic/365718
54. http://www.isnhotnews.com/crime/2009/09/43873
55. http://regist53.blogspot.com/2009/09/blog-post_6448.html
56. http://board.postjung.com/375909.html
57. http://paidoo.net/article/1967789.html
58. http://www.yorkza.com/content/7079/

59. http://phuketenews.blogspot.com/2010/01/posted-by_1450.html
60. http://www.isnhotnews.com/crime/2009/09/45955
61. http://news.ohpicpost.com/%E0%B8%82%E0%B9%88%E0%B8%B2%E0%B8%A7/16466/
62. http://www.ryt9.com/s/bmnd/655512
63. http://www.ceramicdd.com/index.php/article/5-hot-new/1202-2009-11-02-17-14-17
64. http://news.ohpicpost.com/%E0%B8%82%E0%B9%88%E0%B8%B2%E0%B8%A7/16466/
65. http://www.plazazone.com/b20/1407-12402/
66. http://www.click.in.th/news/crime/28307.html
67. http://thairecent.com/Local/2009/251782/
68. http://webboard.yenta4.com/topic/60215
69. http://soi8.forumotion.com/forum-f28/topic-t331.htm
70. http://www.startpage.in.th/view/17636
71. http://www.huaybondin.net/forum/thread-40920-1-1.html
72. http://thailandza.exteen.com/20090219/entry-3
73. http://phuketenews.blogspot.com/2010/01/blog-post_29.html
74. http://www.click.in.th/breaking-news/28409.html
75. http://www.startpage.in.th/view/44834
76. http://onair.kapook.com/notice/64415
77. http://www2.nurnia.com/15421/01/thai-social-political-economic/
78. http://www.opensubscriber.com/message/mass-groups@googlegroups.com/11376881.html
79. http://koreastory.org/philipino-die-because-no-body-song/
80. http://www.readtu.com/content/view/id/101243
81. http://board.siamjung.com/index.php/topic,1249.0.html
82. http://www.ryt9.com/s/bmnd/655770
83. http://news.thaipick.com/news/5373.html
84. http://webboard.yenta4.com/topic/79311
85. http://board.postjung.com/439963.html
86. http://www.ryt9.com/s/bmnd/664713
87. http://www.munnook.com/thread-46086-1-1.html
88. http://atcloud.com/stories/47085

89. http://talk.mthai.com/topic/37286
90. http://www.ryt9.com/s/bmnd/670097
91. http://techkr.com/political-opinion/comstory-14/
92. http://www.click.in.th/breaking-news/45417.html
93. http://www.ohthai.net/15544.html
94. http://thairecent.com/Crime/2010/488894/
95. http://www.ryt9.com/s/bmnd/672770
96. http://www.ryt9.com/s/bmnd/672770
97. http://www.huaybondin.net/forum/thread-55590-1-1.html
98. http://board.postjung.com/443154.html
99. http://talk.mthai.com/topic/33014
100. http://www.ryt9.com/s/bmnd/673182
101. http://board.postjung.com/421197.html
102. http://www.munnook.com/thread-65290-1-6.html
103. http://www.click.in.th/news/crime/38547.html
104. http://www.yorkza.com/content/6985/
105. http://thairecent.com/Local/2009/185767/
106. http://www.fwdder.com/topic/14168
107. http://www.click.in.th/news/around/3829.html
108. http://tnews.teenee.com/crime/17927.html
109. http://www.yorkza.com/content/8183/
110. http://www.ryt9.com/s/bmnd/689983
111. http://atcloud.com/stories/58278
112. http://thairecent.com/First/2010/490180/
113. http://board.postjung.com/430093.html
114. http://www.ryt9.com/s/bmnd/742336
115. http://www.ohthai.net/15552.html
116. http://www.zone-it.com/131792
117. http://www.ohthai.net/15207.html
118. http://www.equinenow.com/video-num-577479.htm

119. http://clipnabber.com/video/Ts8z6xnfv4A.html

120. http://atcloud.com/stories/70515

121. http://www.yorkza.com/content/7079/

122. http://www.ryt9.com/s/bmnd/656134

123. http://tnews.teenee.com/crime/5073.html

124. http://guru.google.co.th/guru/thread?tid=3b488b834e87970f

125. http://board.postjung.com/m/375909.html

126. http://board.postjung.com/375909.html

127. http://play.kapook.com/vdo/show-84258

128. http://talk.mthai.com/topic/31917

129. http://atcloud.com/stories/15667

130. http://www2.nurnia.com/4544/08/thai-social-political-economic/

131. http://www.yorkza.com/content/958/

132. http://talk.mthai.com/topic/44791

133. http://thairecent.com/Local/2010/498064/

134. http://news.sabsan.com/5565.html

135. http://www.ryt9.com/s/bmnd/701777

136. http://www.click.in.th/breaking-news/16079.html

137. http://www.secure2home.com/news/news109.html

138. http://www.hi5thai.com/thread-30097-1-1.html

139. http://www.yorkza.com/content/8787/

140. http://atcloud.com/stories/24562

141. http://www.catchh.com/stories/entry-1490.html

142. http://forums.dp.in.th/thread-1518-1-1.html

143. http://www.click.in.th/news/crime/42712.html

144. http://thairecent.com/Crime/2009/351654/

145. http://webboard.yenta4.com/topic/157821

146. http://www.startpage.in.th/view/49625

147. http://www.munnook.com/thread-49025-1-1.html

148. http://atcloud.com/stories/19885

149. http://regist53.blogspot.com/2009/09/11.html

150. http://board.postjung.com/m/405487.html

151. http://article.wn.com/view/WNAT60daa4f4be70bffc824f7cc31d78f0f4/

152. http://www.yorkza.com/content/2660/

153. http://gotoknow.org/blog/singkhon/224225

154. http://www.siamarchives.com/node/9291

155. http://thairecent.com/Crime/2010/501027/

156. http://talk.mthai.com/topic/26402

157. http://www.yorkza.com/content/2570/

158. http://www.click.in.th/breaking-news/45903.html

159. http://www.childmedia.net/node/507

160. http://learners.in.th/blog/borom7/291997

161. http://paidoo.net/article/546618.html

162. http://seedcom.thai-forum.net/forum-f35/topic-t10616.htm

163. http://tnews.teenee.com/crime/11326.html

164. http://thairecent.com/Local/2010/473775/

165. http://news.ohpicpost.com/%E0%B8%82%E0%B9%88%E0%B8%B2%E0%B8%A7/9379/

166. http://ict.in.th/index.php/topic,3873.0.html

167. http://paidoo.net/article/1903837.html

168. http://forum.siam2fun.com/archiver/?tid-860353.html

169. http://thairecent.com/First/2009/447181/

170. http://www.huaybondin.net/forum/thread-5664-1-1.html

171. http://thairecent.com/First/2010/474919/

172. http://board.postjung.com/407604.html

173. http://www.opensubscriber.com/message/mass-groups@googlegroups.com/10549601.html

174. http://regist53.blogspot.com/2009/04/2_27.html

175. http://board.postjung.com/435945.html

176. http://thairecent.com/Local/2010/475879/

177. http://news.mthai.com/general-news/62539.html

178. http://www.click.in.th/breaking-news/16056.html

179. http://crazymotion.net/-31/wiHT1TA22FkgjAy.html

180. http://www2.nurnia.com/16157/02/thai-social-political-economic/

181. http://www.equinenow.com/video-num-637917.htm

182. http://dekmoram.com/news/hot-news/112-news.html

183. http://www2.nurnia.com/14460/01/thai-social-political-economic/

184. http://www.isnhotnews.com/crime/2009/07/20066

185. http://gotoknow.org/blog/asstudent/62382

186. http://www.pitakthai.com/social/3658.html

187. http://forum.sabyezone.com/index.php/topic,1792.0.html

188. http://zyntag.com/tags/video/_ohoEx15fqg/

189. http://www.pitakthai.com/crime/1145.html

190. http://www.munnook.com/thread-36086-1-9.html

191. http://sexy.yorkza.com/content/4396/

192. http://www.isnhotnews.com/crime/2009/09/45955

193. http://www.pitakthai.com/social/3900.html

194. http://ict.in.th/1666

195. http://www.plazazone.com/b20/2008-62220/

196. http://thairecent.com/Crime/2010/491242/

197. http://www.click.in.th/news/local/692.html

198. http://www.ryt9.com/s/bmnd/688923

199. http://www.newswit.net/read/936351.html

200. http://thairecent.com/First/2010/502150/

ภาคผนวก ง

รายชื่อเว็บไซต์ลามกอนาจาร 150 เว็บไซต์

รายชื่อเว็บไซต์ลามกอนาจาร 150 เว็บไซต์ มีดังนี้

1. http://www.yedke.com/
2. http://thaisex.6x6.in/103/
3. http://dakba.thai-forum.net/forum-f23/topic-t12.htm
4. http://artyz.wapgem.com/11
5. http://ss.comparenotebook.info/lastest/1361.html
6. http://thaig.informe.com/forum/u-uo-u-u-u-ua-u-u-ui-u-a-u-n-u-uo-dt3255.html
7. http://www.thaizexstory.com/home/story/219
8. http://missmovie.forumotion.net/forum-f7/topic-t16.htm
9. http://sport.teenee.com/sport/29985.html
10. http://club.postjung.com/2792-board-17081.html
11. http://www.thaisanook.co.cc/boy/2-boy/4-02620.html
12. http://avzone.wordpress.com/2009/02/27/316/
13. http://nungxonline.findtalk.net/forum-f49/topic-t4.htm
14. http://thaisex.6x6.in/105/
15. http://thaig.informe.com/forum/uo-uu-u-u-u-n-u-u-uo-u-ua-u-uo-u-u-dt719.html
16. http://www.praduk.com/toonx/
17. http://www.zeedasia.com/forums/archiver/tid-14.html
18. http://www.zeedasia.com/forums/thread-10551-1-1.html
19. http://oxz.freesemantic.net/hotest/1133.html
20. http://ss.comparenotebook.info/goURL/4241.html
21. http://www.gnv3.net/bbs/thread-88355-1-1.html
22. http://www.gaykung.com/
23. http://www.thaizexstory.com/home/story/246
24. http://www.thaixxxstory.com/thread-687-1-1.html
25. http://www8.mobileacce.info/last/12727.html
26. http://www.thaizexstory.com/home/story/269
27. http://www8.mobileacce.info/last/12671.html
28. http://bluesea.freesemantic.net/serf/2728.html

29. http://www.g-gang.com/forums/archive/index.php/f-17-p-4.html
30. http://www.saeplay.com/forum-2-1.html
31. http://2pmweb.com/board/
32. http://www.zeedasia.com/forums/thread-8658-1-1.html
33. http://www.gnv3.net/bbs/thread-38255-1-1.html
34. http://ninem.forumotions.com/forum-f3/topic-t4.htm
35. http://seedcom.thai-forum.net/forum-f32/topic-t6019.htm
36. http://bbigbuff.spaces.live.com/blog/
37. http://www.saeplay.com/thread-125-1-1.html
38. http://www.zeedasia.com/forums/thread-11108-1-1.html
39. http://bbs.u18up.com/thread-809-1-5.html
40. http://www.hi5thai.com/archiver/tid-17844.html
41. http://www.saeplay.com/thread-124-1-1.html
42. http://bbs.u18up.com/thread-409-1-9.html
43. http://www.opensubscriber.com/message/saendee-groups@googlegroups.com/13120610.html
44. http://seedcom.thai-forum.net/forum-f32/topic-t7233.htm
45. http://www.thaizexstory.com/home/story/350
46. http://thaisex.6x6.in/27/
47. http://www3.lyricscom.info/lasted_stories/9531.html
48. http://www.surcentro.com/en/info/www.pee-kun.com
49. http://www.g-gang.com/forums/archive/index.php/t-7270.html
50. http://www3.lyricscom.info/lasted_stories/12091.html
51. http://ss.comparenotebook.info/goURL/180.html
52. http://www.zeedasia.com/forums/archiver/tid-13697.html
53. http://www8.mobileacce.info/last/13152.html
54. http://avzone.wordpress.com/2009/01/
55. http://www.thaizexstory.com/home/story/128
56. http://pussyteencam.com/thailand.html
57. http://seedcom.thai-forum.net/forum-f32/topic-t7228.htm

58. http://www.over18x.com/Breakspells-vol-3.xxx
59. http://www.thaizexstory.com/home/story/141
60. http://bluesea.freesemantic.net/serf/519.html
61. http://missmovie.forumotion.net/forum-f7/topic-t101.htm
62. http://ss.comparenotebook.info/goURL/3501.html
63. http://www3.lyricscom.info/lasted_stories/8090.html
64. http://bluesea.freesemantic.net/serf/3000.html
65. http://sexy-porno.org/forum/
66. http://bbs.u18up.com/thread-687-1-1.html
67. http://oxz.freesemantic.net/hotest/3923.html
68. http://thaisex.6x6.in/page/5/
69. http://seedcom.thai-forum.net/forum-f32/topic-t4767.htm
70. http://artyz.wapgem.com/4
71. http://seed.gameref.info/go/94.html
72. http://www.zeedasia.com/forums/archiver/tid-11067.html
73. http://www.g-gang.com/forums/archive/index.php/t-7272.html
74. http://www.thaizexstory.com/home/story/118
75. http://www.opensubscriber.com/message/saendee-groups@googlegroups.com/13012284.html
76. http://oxz.freesemantic.net/serf/2529.html
77. http://www.zeedasia.com/forums/archiver/tid-8241.html
78. http://planet.kapook.com/sexyboy2/blog/viewnew/68996
79. http://www.zeedasia.com/forums/archiver/tid-13119.html
80. http://forum.soda-zaa.com/thread-5778-1-2.html
81. http://space.postjung.com/1083242-blog-62869.html
82. http://oxz.freesemantic.net/serf/68.html
83. http://www.saeplay.com/thread-31-1-1.html
84. http://ss.comparenotebook.info/lastest/2834.html
85. http://www.thaizexstory.com/home/story/188
86. http://www8.mobileacce.info/last/15090.html

87. http://forum.soda-zaa.com/thread-2514-1-3.html
88. http://seed.gameref.info/go/4139.html
89. http://seedcom.thai-forum.net/forum-f32/topic-t8555.htm
90. http://missmovie.forumotion.net/forum-f7/topic-t101-15.htm
91. http://bbs.u18up.com/thread-1593-1-1.html
92. http://seed.gameref.info/go/1709.html
93. http://swing.realbb.net/forum-f1/topic-t1.htm
94. http://www.opensubscriber.com/message/saendee-groups@googlegroups.com/13098555.html
95. http://seedcom.thai-forum.net/forum-f32/topic-t7491.htm
96. http://oxz.freesemantic.net/hotest/2055.html
97. http://forum.soda-zaa.com/thread-2526-1-2.html
98. http://forum.soda-zaa.com/thread-2460-1-5.html
99. http://www.sexy-picpost-xxx.com/
100. http://www3.lyricscom.info/lasted_stories/10766.html
101. http://missmovie.forumotion.net/forum-f7/topic-t99.htm
102. http://seed.gameref.info/hs/3330.html
103. http://www.zeedasia.com/forums/thread-13828-1-4.html
104. http://www.saeplay.com/viewthread.php?tid=65
105. http://www.69rental.com/
106. http://forum.soda-zaa.com/thread-6351-1-5.html
107. http://bbs.u18up.com/thread-827-1-1.html
108. http://www.thaizexstory.com/home/story/160
109. http://thaisex.6x6.in/69/
110. http://www.peekstats.com/www.gaykung.com
111. http://www.peekstats.com/www.ohjeed.com
112. http://www3.lyricscom.info/lasted_stories/9091.html
113. http://yedke.wordpress.com/2008/12/
114. http://www.thaixxxstory.com/thread-2582-1-1.html
115. http://www.ohpicpost.com/story/thread-40485-1-5.html

116. http://x691.com/forum/index.php

117. http://seedcom.thai-forum.net/forum-f32/topic-t7811.htm

118. http://www.thaizexstory.com/home/story/94

119. http://www.giggay.com/board/friend/topic-2192

120. http://thaix.in/thread-16317-1-4.html

121. http://seedcom.thai-forum.net/forum-f32/topic-t5989.htm

122. http://forum.soda-zaa.com/thread-2521-1-3.html

123. http://www.zeedasia.com/forums/archiver/tid-1540.html

124. http://oxz.freesemantic.net/serf/2763.html

125. http://www.thaizexstory.com/home/story/217

126. http://www8.mobileacce.info/last/13066.html

127. http://www.saeplay.com/thread-127-1-1.html

128. http://bluesea.freesemantic.net/hotest/3465.html

129. http://seedcom.thai-forum.net/forum-f32/topic-t4562.htm

130. http://atcloud.com/stories/49688

131. http://www.praduk.com/toonx/tag/

132. http://www.dek-it.net/read-htm-tid-8875.html

133. http://www.pakkdee.info/board/

134. http://bbs.u18up.com/thread-1025-1-1.html

135. http://www.googgang.com/clip/

136. http://xstory.aoi2you.com/

137. http://www.surcentro.com/en/info/www.clipxx.com

138. http://freeclippo.com/board/

139. http://thaix.in/

140. http://bbs.powerzeed.com/

141. http://seed.gameref.info/go/1190.html

142. http://www2.lhc-effect.info/

143. http://www.yedme.com/

144. http://www.fwdder.com/topic/4679

145. http://thaig.informe.com/forum/message45525.html

146. http://board.darkconceal.com/
147. http://goolabua.com/
148. http://www.goox.info/
149. http://picpost.gig01.com/
150. http://www.cdxyz.net/

ภาคผนวก จ

รายชื่อเว็บไซต์ปกติ 200 เว็บไซต์

รายชื่อเว็บไซต์ปกติ 200 เว็บไซต์ มีดังนี้

1. http://www.sansuk.com
2. http://www.108event.com
3. http://www.clinicrak.com
4. http://www.sodamag.net
5. http://www.Exteen.com
6. http://www.Ohozaa.com
7. http://www.Meemodel.com
8. http://www.madoo.com
9. http://www.yenta4.om
10. http://www.siamzone.com
11. http://www.tlcthai.com
12. http://www.tarad.com
13. http://www.siamha.com
14. http://www.365jukebox.com
15. http://www.zuzaa.com
16. http://www.jikgo.com
17. http://www.siamdara.com
18. http://www.zubzip.com
19. http://www.yimza.com
20. http://www.seezaa.com
21. http://www.jkdramas.com
22. http://www.108ideas.com
23. http://www.sbuyzone.com
24. http://www.narak.com
25. http://www.nurnia.com
26. http://www.zheza.com
27. http://www.samarts.com
28. http://www.rannaidee.com

29. http://coolzshot.com
30. http://www.khaikai.com
31. http://www.bkkonline.com
32. http://www.siam2you.com
33. http://www.impressivezone.com
34. http://www.aonlines.com
35. http://www.sabyedigg.com
36. http://www.oopza.com
37. http://www.siamtrue.com
38. http://www.namtea.com
39. http://www.diggma.com
40. http://www.zabjung.com
41. http://www.moomdigg.com
42. http://www.eduzones.com
43. http://www.tourthai.com
44. http://www.skydigg.com
45. http://www.haarai.com
46. http://www.discuzthai.com
47. http://www.medias.co.th
48. http://www.cmsthailand.com
49. http://movie.classifiedthai.com
50. http://www.arunsawat.com
51. http://www.rsu-cyberu.com
52. http://www.krusu.com
53. http://www.moe.go.th/idea
54. http://graduate.kru.ac.th/course
55. http://www.chaiwbi.com
56. http://www.crnfe.ac.th
57. http://www.rjanadd.com
58. http://www.edutoday.in.th

59. http://www.niyada.net
60. http://www.learnland.net
61. http://www.kanngan.com
62. http://www.krusupap.com
63. http://www.taklong.com
64. http://www.thecameracity.com
65. http://www.rpst.or.th
66. http://www.pixprox.net
67. http://www.pixnice.com
68. http://www.photothai.net
69. http://www.artphotoschool.com/
70. http://www.ohophoto.com
71. http://www.rpst-digital.org
72. http://www.thairetouch.com
73. http://www.hardcoregraphic.com
74. http://www.thaigraph.com
75. http://www.rookienet.com
76. http://www.designparty.com
77. http://www.thai3d.net
78. http://www.triamudom.ac.th
79. http://www2.bodin.ac.th
80. http://www.satriwit.ac.th
81. http://www.snr.ac.th
82. http://www.sk.ac.th
83. http://www.phrachaokrungthon.com
84. http://www.chartthai.or.th
85. http://www.archae.go.th
86. http://www.kanmuang.org
87. http://www.democrat.or.th
88. http://www.ubudpa.in.th

89. http://www.pol.ru.ac.th
90. http://www.terdchai.com
91. http://www.krooadd.com
92. http://www.ftawatch.org
93. http://www.kru-itth.com
94. http://www.trainer.in.th
95. http://www.metukyang.com
96. http://www.stks.or.th
97. http://www.tkpark.or.th
98. http://www.radompon.com
99. http://www.tutoronline.co.th
100. http://www.ecommerce.or.th
101. http://www.sanook.com
102. http://www.kapook.com
103. http://www.mthai.com
104. http://www.manager.co.th
105. http://www.dek-d.com
106. http://www.teenee.com
107. http://www.exteen.com
108. http://www.bloggang.com
109. http://www.playpark.com
110. http://www.narak.com
111. http://www.meemodel.com
112. http://www.asiasoft.co.th
113. http://www.yenta4.com
114. http://www.postjung.com
115. http://www.siamzone.com
116. http://www.siamsport.co.th
117. http://www.gg.in.th
118. http://www.tlcthai.com

119. http://www.212cafe.com
120. http://www.pramool.com
121. http://www.siamza.com
122. http://www.thaicybergames.com
123. http://www.hunsa.com
124. http://www.uploadtoday.com
125. http://www.truelife.com
126. http://www.siamha.com
127. http://www.ini3.co.th
128. http://www.soccersuck.com
129. http://www.pantipmarket.com
130. http://www.ohozaa.com
131. http://www.tarad.com
132. http://www.zheza.com
133. http://www.thaiSecondhand.com
134. http://www.one2car.com
135. http://www.ob.tc
136. http://www.uppic.net
137. http://www.thaiza.com
138. http://www.dailynews.co.th
139. http://www.gmember.com
140. http://www.thaiware.com
141. http://www.temppic.com
142. http://www.gushare.com
143. http://www.popcornfor2.com
144. http://www.adintrend.com
145. http://www.siamphone.com
146. http://www.clipmass.com
147. http://www.thaimisc.com
148. http://www.bitthailand.com

149. http://www.siambit.com

150. http://www.math26.com

151. http://www.cuas.or.th

152. http://www.niets.or.th

153. http://www.vcharkarn.com/vexam

154. http://www.campus.sanook.com/entrance

155. http://www.admissions.chula.ac.th

156. http://www.yenta4.com/entrance

157. http://www.pantip.com/tech

158. http://www.bcoms.net

159. http://www.com-th.net

160. http://www.notebookspec.com

161. http://www.adslthailand.com

162. http://www.unlimitpc.com

163. http://www.it-guides.com

164. http://www.divland.com

165. http://www.thaiopensource.org

166. http://www.opentle.org

167. http://www.thaimsn.net

168. http://www.mrpalm.com

169. http://www.monavista.com

170. http://www.mac2hand.com

171. http://www.hi5thai.com

172. http://www.thaiall.com

173. http://www.thaiiphoneclub.com

174. http://www.studentloan.ktb.co.th

175. http://www.studentloan.or.th

176. http://www.ex-mba.com

177. http://www.ocsc.go.th

178. http://www.ies-education.com

179. http://www.educatepark.com

180. http://www.yesthailand.org

181. http://www.it-ed.com

182. http://www.coj.go.th

183. http://www.moj.go.th

184. http://www.admincourt.go.th

185. http://www.lawyerthai.com

186. http://www.dhamma.th.gs

187. http://www.luangta.com

188. http://www.dhammathai.org

189. http://www.sdsweb.org

190. http://www.catholic.or.th

191. http://www.thaibible.net

192. http://www.kondee.com

193. http://www.netdesign.ac.th

194. http://www.artanddesign.ac.th

195. http://www.allstep.net

196. http://www.3d-dsign.com

197. http://www.aptech.co.th

198. http://www.ecc.ac.th

199. http://www.greatfriends.biz

200. http://www.computer.ru.ac.th

ภาคผนวก ฉ

กลุ่มคำลามกอนาจาร 112 คำ

กลุ่มคำลามกอนาจาร 112 คำ

1. คำอนาจารที่ 1	ซี๊ด
2. คำอนาจารที่ 2	ถ่างขา
3. คำอนาจารที่ 3	กระเจี๊ยว
4. คำอนาจารที่ 4	เสียตัว
5. คำอนาจารที่ 5	แทงรู
6. คำอนาจารที่ 6	เอากัน
7. คำอนาจารที่ 7	หนังเอ็กซ์
8. คำอนาจารที่ 8	แอบดู
9. คำอนาจารที่ 9	เสียวดี
10. คำอนาจารที่ 10	ดาราโป๊
11. คำอนาจารที่ 11	การ์ตูนโป๊
12. คำอนาจารที่ 12	แก้ผ้า
13. คำอนาจารที่ 13	ภาพหลุด
14. คำอนาจารที่ 14	เย็ด
15. คำอนาจารที่ 15	แลกลิ้น
16. คำอนาจารที่ 16	สยิว
17. คำอนาจารที่ 17	หนังโป๊
18. คำอนาจารที่ 18	ประสบการณ์เสียว
19. คำอนาจารที่ 19	ปิดบริสุทธิ์
20. คำอนาจารที่ 20	ปลื้องผ้า
21. คำอนาจารที่ 21	เล่นเสียว
22. คำอนาจารที่ 22	เสียวจัง
23. คำอนาจารที่ 23	เสียวหี
24. คำอนาจารที่ 24	ดูดนม
25. คำอนาจารที่ 25	สวิงกิ้ง
26. คำอนาจารที่ 26	ชักว่าว
27. คำอนาจารที่ 27	เลียหี
28. คำอนาจารที่ 28	เสียวหอย

29. คำอนาจารที่ 29 เลียหอย
30. คำอนาจารที่ 30 ลีลารัก
31. คำอนาจารที่ 31 รูปเกย์
32. คำอนาจารที่ 32 เอาหอย
33. คำอนาจารที่ 33 คลิปลับ
34. คำอนาจารที่ 34 แอบถ่ายน้อง
35. คำอนาจารที่ 35 อึ๊บ
36. คำอนาจารที่ 36 ท่าร่วมเพศ
37. คำอนาจารที่ 37 เซ็กซ์
38. คำอนาจารที่ 38 คลิปมือถือ
39. คำอนาจารที่ 39 แอบถ่าย
40. คำอนาจารที่ 40 ภาพโป๊
41. คำอนาจารที่ 41 เรื่องเสียว
42. คำอนาจารที่ 42 จุดเสียว
43. คำอนาจารที่ 43 หัวนม
44. คำอนาจารที่ 44 เลียนม
45. คำอนาจารที่ 45 น้ำแตก
46. คำอนาจารที่ 46 เสียงคราง
47. คำอนาจารที่ 47 อมควย
48. คำอนาจารที่ 48 บีบนม
49. คำอนาจารที่ 49 เนินหอย
50. คำอนาจารที่ 50 ซอยถี่
51. คำอนาจารที่ 51 ภาพเสียว
52. คำอนาจารที่ 52 ภาพหลุดดารา
53. คำอนาจารที่ 53 รูปแอบถ่าย
54. คำอนาจารที่ 54 สื่อลามก
55. คำอนาจารที่ 55 นมหก
56. คำอนาจารที่ 56 ร่วมเพศ
57. คำอนาจารที่ 57 ร่วมรัก
58. คำอนาจารที่ 58 จุดสุดยอด

59. คำอนาจารที่ 59 เสียสาว
60. คำอนาจารที่ 60 น้ำรัก
61. คำอนาจารที่ 61 จิ๋ม
62. คำอนาจารที่ 62 คลิปหลุด
63. คำอนาจารที่ 63 เงี่ยน
64. คำอนาจารที่ 64 เร่าร้อน
65. คำอนาจารที่ 65 นมขนาดใหญ่
66. คำอนาจารที่ 66 อมเต็มปาก
67. คำอนาจารที่ 67 เสียวซ่าน
68. คำอนาจารที่ 68 รูดเขารูดออก
69. คำอนาจารที่ 69 เป้ากางเกง
70. คำอนาจารที่ 70 ไซร้ซอกคอ
71. คำอนาจารที่ 71 ปลดผ้าขนหนู
72. คำอนาจารที่ 72 ถอดกางเกง
73. คำอนาจารที่ 73 ขึ้นคร่อม
74. คำอนาจารที่ 74 ท่อนจรวด
75. คำอนาจารที่ 75 ช่วยตัวเอง
76. คำอนาจารที่ 76 เปลี่ยนคู่
77. คำอนาจารที่ 77 ปลุกอารมณ์
78. คำอนาจารที่ 78 กระหรี่
79. คำอนาจารที่ 79 ถ่างแข้งถ่างขา
80. คำอนาจารที่ 80 ถอดเสื้อผ้า
81. คำอนาจารที่ 81 อวบอิ่ม
82. คำอนาจารที่ 82 ปลดกระดุม
83. คำอนาจารที่ 83 นอนถ่างขา
84. คำอนาจารที่ 84 รูปเปลือย
85. คำอนาจารที่ 85 กล่องดวงใจ
86. คำอนาจารที่ 86 ล่อลื่น
87. คำอนาจารที่ 87 ถูไถ
88. คำอนาจารที่ 88 เล่นว่าว

89. คำอนาจารที่ 89 ดูดปาก
90. คำอนาจารที่ 90 ทะลวงรู
91. คำอนาจารที่ 91 แข็งโด่
92. คำอนาจารที่ 92 กระตุกเกร็ง
93. คำอนาจารที่ 93 กระแทกซอย
94. คำอนาจารที่ 94 แก้มก้น
95. คำอนาจารที่ 95 ท่อนเอ็น
96. คำอนาจารที่ 96 เสียวงูบ
97. คำอนาจารที่ 97 ยัดพรวด
98. คำอนาจารที่ 98 ถลกกระโปรง
99. คำอนาจารที่ 99 โก้งโค้ง
100. คำอนาจารที่ 100 ติ่งแตด
101. คำอนาจารที่ 101 ร้อนวูบวาบ
102. คำอนาจารที่ 102 กอดรัด
103. คำอนาจารที่ 103 ร้องครวญคราง
104. คำอนาจารที่ 104 ร่างเปลือย
105. คำอนาจารที่ 105 สอดใส่
106. คำอนาจารที่ 106 อ้าขา
107. คำอนาจารที่ 107 มิดด้าม
108. คำอนาจารที่ 108 แก้ผ้า
109. คำอนาจารที่ 109 กระแทกแรง
110. คำอนาจารที่ 110 ลักหลับ
111. คำอนาจารที่ 111 บดจูบ
112. คำอนาจารที่ 112 น้ำล่อลื่น

ประวัติผู้วิจัย

ชื่อ - สกุล ชาญพัฒณ์ ภินันท์รัชต์ธร

ที่อยู่ 219 หมู่8 ต.กำแพงแสน อ.กำแพงแสน จ.นครปฐม 73140

ที่ทำงาน 55-57-59 ตรอกต้นโพธิ์ ต.พระปฐมเจดีย์ อ.เมือง จ.นครปฐม 73000 โทรศัพท์ (034) 250962-4

ประวัติการศึกษา

พ.ศ. 2547	สำเร็จการศึกษาปริญญาวิทยาศาสตร์บัณฑิต วิชาเอกคณิตศาสตร์ จาก คณะวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี
พ.ศ. 2548	ศึกษาต่อระดับปริญญามหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ประวัติการทำงาน

พ.ศ. 2548 – ปัจจุบัน	ผู้ดูแลระบบ บริษัทสโมทัยคอมพิวเตอร์