ระบบตรวจหาข้อมูลอันตรายในเวิร์ลไวด์เว็บ

WORLD WIDE WEB MALEFIC CODE INSPECTION SYSTEM

นางสาว พิชานี ทัศนเสถียร

นาย ภักดิ์ทูล ใจทอง

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2549

ระบบตรวจหาข้อมูลอันตรายในเวิร์ลไวด์เว็บ

WORLD WIDE WEB MALEFIC CODE INSPECTION SYSTEM

โดย

นางสาว พิชานี ทัศนเสถียร

นาย ภักดิ์ทูล ใจทอง

อาจารย์ที่ปรึกษา

อาจารย์ ธนัญชัย ตรีภาค

อาจารย์ อัครเดช วัชระภูพงษ์

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2549

ปริญญานิพนธ์ปีการศึกษา 2549

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง ระบบตรวจหาข้อมูลอันตรายในเวิร์ลไวด์เว็บ

WORLD WIDE WEB MALEFIC CODE INSPECTION SYSTEM

ผู้จัดทำ

1. นางสาว พิชานี ทัศนเสถียร รหัสประจำตัว 46010522
2. นาย ภักดิ์ทูล ใจทอง รหัสประจำตัว 46010555

_____อาจารย์ที่ปรึกษา

(อาจารย์ ธนัญชัย ตรีภาค)

_____อาจารย์ที่ปรึกษา

(อาจารย์ อัครเดช วัชระภูพงษ์)

ระบบตรวจหาข้อมูลอันตรายในเวิร์ลไวด์เว็บ

นางสาว พิชานี ทัศนเสถียร	46010522
นาย ภักดิ์ทูล ใจทอง	46010555
อาจารย์ ธนัญชัย ตรีภาค	อาจารย์ที่ปรึกษา
อาจารย์ อัครเดช วัชระภูพงษ์	อาจารย์ที่ปรึกษาร่วม
ปีการศึกษา 2549	

บทคัดย่อ

เนื่องด้วยในปัจจุบันภาครัฐบาลมีนโยบายในการป้องกันและปราบปรามเว็บไซต์ที่มีเนื้อหาไม่เหมาะสม โดยกระบวนการตรวจสอบเว็บไซต์ดังกล่าวดำเนินการด้วยวิธีการตรวจสอบด้วยทีมงานของหน่วยงานราชการที่รับผิดชอบ ซึ่งล่าช้าและไม่ทันต่อการเพิ่มจำนวนของเว็บไซต์

โครงงานนี้จึงได้จัดทำขึ้นเพื่อให้สามารถค้นหาเว็บไซต์ที่มีเนื้อหาไม่เหมาะสมให้มีประสิทธิภาพมากยิ่งขึ้น โดยสร้างระบบอัตโนมัติที่สามารถค้นหารายการเว็บที่มีเนื้อหาไม่เหมาะสมในอินเตอร์เน็ตอยู่ตลอดเวลา ในการตรวจสอบใช้การวิเคราะห์รูปภาพโดยใช้หลักของการประมวลผลรูปภาพ และเทคนิคการวิเคราะห์เนื้อหาในเว็บไซต์ เมื่อผ่านกระบวนการทั้งหมดจะทำการวิเคราะห์หาค่าน้ำหนักของความไม่เหมาะสมของเว็บไซต์นั้นๆว่ามีค่าความไม่เหมาะสมมากน้อยเพียงใด ถ้าเว็บไซต์ใดถูกวิเคราะห์ว่าเป็นเว็บไซต์ที่มีเนื้อหาไม่เหมาะสมจะถูกเก็บไว้ในรายการเว็บไซต์ต้องห้ามเพื่อให้ภาครัฐบาลดำเนินการกับเว็บไซต์นั้นๆ ในขั้นตอนต่อไป

World Wide Web Malefic Code Inspection System

Ms. Pichanee Tassanasatien	46010522
Mr. Paktoon Jaithong	46010555
Mr. Thananchai Treepak	Advisor
Mr. Akkradach Watcharapupong	Co-Advisor

Academic Year 2006

ABSTRACT

Nowadays, Ministry of information and communication technology (ICT) proposes strict policy to prevent and suppress Thai people access any pornographic and inappropriate website. Group of authorized government officers named "Cyber Inspector" was established to investigate and block any website which contains pornographic content from all internet users. By using labor-intensive detection method, Cyber Inspector group unsuccessful in battle with porn websites because growth rate of porn websites is much more than human can identify manually.

For this reason, our project proposes an automatic pornographic website discovery system that recursive crawling websites in the internet and verifies that they are pornographic websites. In verifying process, our system uses not only image processing technique but also text analysis method to extract information from websites and then employ principal of component analysis to make a decision that each website is porn website or not. For all of pornographic website we get from our system, we collect them into database for future ICT work out.

กิตติกรรมประกาศ

ปริญญานิพนธ์ฉบับนี้สำเร็จลุล่วงได้อย่างดีด้วยคำแนะนำและคำปรึกษาเกี่ยวกับการดำเนินการศึกษาและวิจัยจากอาจารย์ธนัญชัย ตรีภาค ซึ่งเป็นอาจารย์ผู้ควบคุมปริญญานิพนธ์และอาจารย์อัครเดช วัชระภูพงษ์ ผู้ควบคุมปริญญานิพนธ์ร่วม ผู้วิจัยรู้สึกซาบซึ้งในความอนุเคราะห์จากท่านอาจารย์ทั้งสองและขอกราบขอบพระคุณเป็นอย่างสูง

ขอขอบคุณห้องวิจัย ISAG ภาควิชาวิศวกรรมคอมพิวเตอร์ ที่ได้สนับสนุนเครื่องมือตลอดจนข้อมูล และหนังสือต่างๆที่ใช้ในการทำวิจัย

ขอกราบขอบพระคุณคณาจารย์ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุกๆท่านที่ได้ประสิทธิ์ประสาทวิชาให้กับข้าพเจ้า

ขอขอบคุณเพื่อน ๆ พี่ ๆ น้องๆ ในภาควิชาวิศวกรรมคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุกคนที่ให้คำแนะนำต่างๆ และคอยให้กำลังใจเสมอมา

สุดท้ายนี้ข้าพเจ้าขอกราบขอบพระคุณ บิดา มารดา และครอบครัวของข้าพเจ้าที่เป็นกำลังใจ และให้การสนับสนุนในทุกๆเรื่อง ทำให้ข้าพเจ้าสามารถทำปริญญานิพนธ์ฉบับนี้สำเร็จลุล่วงด้วยดี

คุณค่าและประโยชน์อันพึงมาจากปริญญานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบแด่ผู้มีพระคุณทุกท่าน

นางสาว พิชานี ทัศนเสถียร

นาย ภักดิ์ทูล ใจทอง

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

สารบัญรูป (ต่อ)

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของโครงงาน

เทคโนโลยีเครือข่ายและอินเตอร์เน็ทเติบโตขึ้นอย่างรวดเร็ว และเวิร์ลไวด์เว็บได้กลายมาเป็นแหล่งข้อมูลสำคัญ และการสร้างเว็บไซต์สามารถทำได้ง่ายขึ้นประกอบกับผู้มีความรู้ความสนใจทางด้านไอทีมีจำนวนมากยิ่งขึ้น จึงก่อให้เกิดจำนวนเว็บไซต์มากขึ้นในทุกๆวัน ซึ่งเว็บไซต์เหล่านั้นมีทั้งเว็บไซต์ที่มีประโยชน์ให้ความรู้ และเว็บไซต์อันตราย โดยในปัจจุบันวิธีการกลั่นกรองเนื้อหาทางอินเตอร์เน็ทมีหลายวิธีการ สามารถแบ่งออกเป็น 2 กลุ่มวิธีคือ ใช้การกำหนดคำสำคัญ และใช้การพิจารณาด้วยบุคคล ซึ่งวิธีการดังกล่าวไม่มีประสิทธิภาพเพียงพอกับปริมาณเว็บไซต์อันตรายที่เพิ่มขึ้นในแต่ละวัน ดังนั้นจึงควรมีกระบวนการค้นหาและกลั่นกรองเว็บไซต์อันตรายที่มีประสิทธิภาพมากยิ่งขึ้น

1.2 วัตถุประสงค์ของโครงงาน

1. เพื่อเพิ่มประสิทธิภาพในการค้นหาเว็บไซต์อนาจารที่เพิ่มขึ้นอย่างรวดเร็วในแต่ละวัน
2. เพื่อสร้างโปรแกรมที่จะเก็บรวบรวมลิงค์ต่างๆในแต่ละเว็บไซต์เพื่อนำมาหาค่าน้ำหนักความอันตรายของเว็บไซต์นั้น
3. เพื่อสร้างโปรแกรมที่จะสามารถประมวลผลภาพที่อยู่ในแต่ละหน้าเว็บเพจว่าเป็นภาพอนาจารหรือไม่ โดยใช้หลักการของการประมวลผลภาพเข้ามาช่วย
4. เพื่อสร้างโปรแกรมที่จะตรวจสอบหาคำสำคัญที่กำหนดไว้ในหน้าเว็บไซต์หนึ่งๆ ซึ่งนำมาช่วยในการหาน้ำหนักความอันตรายของเว็บไซต์นั้นๆ
5. เพื่อตรวจหารายชื่อเว็บไซต์ที่เข้าข่ายอันตรายและรวบรวมรายชื่อเว็บไซต์ไว้ในฐานข้อมูล

1.3 ขอบเขตของโครงงาน

1. พัฒนาส่วนของการค้นหาลิงค์ต่างๆในหน้าเว็บไซต์โดยใช้เว็บเสิร์จเอนจิ้น
2. พัฒนาส่วนของการวิเคราะห์รูปภาพอนาจารโดยใช้กระบวนการของการประมวลผลรูปภาพซึ่งสามารถวิเคราะห์ภาพที่มีช่วงสีในช่วงที่กำหนดเท่านั้นและภาพต้องอยู่ในสภาวะแสงที่ค่อนข้างคงที่ ไม่มีการเปลี่ยนแปลงมากเกินไป
3. พัฒนาส่วนของการวิเคราะห์เนื้อหาในเว็บไซต์ ซึ่งสามารถวิเคราะห์เนื้อหาที่ใช้ภาษาอังกฤษเป็นหลัก

4. การประมวลข้อมูลและบันทึกลงฐานข้อมูลเพื่อระบุถึงเว็บไซต์ที่มีเนื้อหาไม่เหมาะสม โดยการคำนวณหาค่าน้ำหนักของความไม่เหมาะสมของเว็บไซต์นั้นๆและบันทึกลงฐานข้อมูลจัดทำรายการเว็บไซต์ต้องห้าม
5. พัฒนาส่วนของหน้าแสดงผลรายการเว็บไซต์ที่มีเนื้อหาไม่เหมาะสม

1.4 วิธีการดำเนินการ

1. ศึกษาลักษณะการทำงานของโปรแกรมรวบรวมเอกสารเว็บ
2. ศึกษากระบวนการการประมวลผลรูปภาพ
3. ศึกษากระบวนการวิเคราะห์เนื้อหาภายในเว็บไซต์
4. ศึกษาเกี่ยวกับการออกแบบและการใช้งานฐานข้อมูล
5. ศึกษาการจัดทำเว็บไซต์แสดงผลที่ได้จากระบบ
6. พัฒนาโปรแกรมรวบรวมเอกสารเว็บร่วมกับฐานข้อมูล
7. พัฒนาโปรแกรมการตรวจสอบภาพอนาจาร
8. พัฒนาโปรแกรมการวิเคราะห์เนื้อหาภายในเว็บไซต์ โดยพิจารณาหาคำสำคัญภายในเว็บไซต์
9. ออกแบบฐานข้อมูลเพื่อรองรับการทำงาน
10. พัฒนาโปรแกรมการวิเคราะห์หาค่าน้ำหนักของความไม่เหมาะสมของเว็บไซต์ โดยนำโปรแกรมทั้งในส่วนของการวิเคราะห์ภาพและเนื้อหาประกอบกัน
11. จัดทำเว็บไซต์เพื่อแสดงผลของโปรแกรมเพื่อให้ผู้ใช้สามารถเรียกดูรายการเว็บไซต์ต้องห้ามได้
12. ทดสอบการใช้งานของโปรแกรมที่พัฒนาขึ้นและแก้ไขโปรแกรมให้มีประสิทธิภาพมากยิ่งขึ้น

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้รับความรู้และเข้าใจหลักการทำงานของระบบเสิร์จเอนจิ้น
2. ได้รับความรู้ความเข้าใจเกี่ยวกับการวิเคราะห์รูปภาพที่ประกอบด้วยสีผิวของมนุษย์
3. โปรแกรมที่ช่วยหาเว็บไซต์ที่มีเนื้อหาไม่เหมาะสม
4. โปรแกรมที่ช่วยเพิ่มประสิทธิภาพการค้นหาเว็บไซต์ที่มีเนื้อหาไม่เหมาะสม
5. โปรแกรมที่ช่วยประมวลผลภาพในแต่ละเว็บเพจและวิเคราะห์ความอันตรายของเว็บเพจนั้นให้โดยอัตโนมัติ
6. ส่วนของการแสดงผลเพื่อให้ผู้ใช้สามารถดูรายการเว็บไซต์ต้องห้ามได้

1.6 ส่วนประกอบของปริญญานิพนธ์

ปริญญานิพนธ์ฉบับนี้ได้แบ่งเนื้อหาออกเป็น 5 บทด้วยกันคือ

บทที่ 1 กล่าวถึงความสำคัญและที่มาของโครงงาน วัตถุประสงค์ของโครงงาน ขอบเขตของโครงงาน วิธีการดำเนินการ ประโยชน์ที่คาดว่าจะได้รับ และส่วนประกอบของปริญญานิพนธ์

บทที่ 2 กล่าวถึงทฤษฎีที่เกี่ยวข้องซึ่งจะประกอบไปด้วย ทฤษฎีของโปรแกรมรวบรวมเอกสารเว็บ การประมวลผลรูปภาพ การวิเคราะห์เนื้อหาภายในเว็บไซต์ และการจัดกลุ่มข้อมูล

บทที่ 3 กล่าวถึงการออกแบบและพัฒนา ซึ่งเป็นการออกแบบโครงสร้างของโครงงานและโปรแกรมในแต่ละส่วน

บทที่ 4 กล่าวถึงการทดลองและผลการทดลอง โดยได้เขียนวิธีการทดลองและผลการทดลองของโปรแกรมในแต่ละส่วน

บทที่ 5 เป็นบทวิจารณ์และสรุป ซึ่งกล่าวถึงบทสรุปของโครงงาน วิจารณ์สิ่งที่ได้รับจากโครงงาน และข้อเสนอแนะสำหรับเป็นแนวทางในการพัฒนาต่อ

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

การค้นหาเว็บไซต์ที่มีเนื้อหาไม่เหมาะสมด้วยระบบเสิร์จเอนจิ้นแบบหนึ่งที่เรียกว่า โปรแกรมรวบรวมเอกสารเว็บ (Spider หรือ Crawler) ซึ่งสามารถนำไปใช้งานได้ดีเพราะโปรแกรมรวบรวมเอกสารเว็บจะทำการเข้าไปสำรวจและอ่านเว็บเพจ หากพบลิงค์ก็จะทำการติดตามลิงค์ภายในเว็บไซต์จนครบ เมื่อได้ลิงค์ของเว็บเพจและลิงค์ของภาพแล้วจะนำมาวิเคราะห์คุณสมบัติของภาพ โดยจะใช้หลักการของการประมวลผลภาพ (Image Processing) เข้ามาช่วยเพื่อระบุว่าภาพนั้นเป็นภาพอนาจารหรือไม่ รวมทั้งวิเคราะห์เนื้อหาในเอกสารเว็บ ด้วยคำสำคัญและจำแนกประเภทของเอกสารเว็บนั้นว่ามีความไม่เหมาะสมมากน้อยเพียงใด

2.1 กลไกการทำงานของระบบเสิร์จเอนจิ้น

เสิร์จเอนจิ้นทำงานโดยการเก็บข้อมูลเกี่ยวกับหน้าเว็บเพจ เสิร์จเอนจิ้นจะทำหน้าที่ในการดึงข้อมูลของเว็บเพจแต่ละหน้า และทำการวิเคราะห์ข้อมูลเก็บไว้ในฐานข้อมูลเพื่อไว้ค้นหาในภายหลัง เมื่อผู้ใช้ทำการใช้งานเสิร์จเอนจิ้นและทำการถามหรือค้นหา ตัวอย่างเช่น การค้นหาโดยการใช้คำสำคัญ เสิร์จเอนจิ้นจะทำการมองหาดัชนีเพื่อจัดเตรียมรายชื่อของหน้าเว็บเพจที่ตรงกับความต้องการมากที่สุด โดยใช้กฎเกณฑ์ของเสิร์จเอนจิ้น แต่ก็มีเสิร์จเอนจิ้นประเภทอื่นที่ไม่มีการเก็บดัชนี ข้อมูลที่เสิร์จเอนจิ้นต้องการนั้นใช้เมื่อเกิดการถามขึ้นเท่านั้น ซึ่งจะทำให้ได้ข้อมูลล่าสุดแต่มีข้อเสียคือใช้เวลาในการค้นหานานกว่า โดยสามารถแบ่งกระบวนการทำงานของเสิร์จเอนจิ้นออกเป็น 3 ส่วน

1) โปรแกรมรวบรวมเอกสารเว็บ (Spider หรือ Crawler)
2) รายการดัชนีข้อมูล (Index หรือ Catalog)
3) โปรแกรมการสืบค้น (Search engine software)

ซึ่งในปัจจุบันนี้ได้มีการประยุกต์ใช้งานเสิร์จเอนจิ้นมากมาย เช่น Google, AltaVista, Lycos, Yahoo และ Hotbot เป็นต้น

2.1.1 โปรแกรมรวบรวมเอกสารเว็บ (Spider หรือ Crawler)

โปรแกรมรวบรวมเอกสารเว็บ จะทำหน้าที่สำรวจเว็บจากโดเมนต่าง ๆ โดยโปรแกรมรวบรวมเอกสารเว็บจะเข้าไปสำรวจและอ่านเว็บเพจ และหากพบลิงค์ก็จะทำการติดตามลิงค์ภายในเว็บไซต์จนครบ ซึ่งจากการทำงานในลักษณะโยงใยนี้ จึงเป็น ที่มาของคำว่า Spider หรือ

Crawler จากนั้นโปรแกรมรวบรวมเอกสารเว็บจะนำข้อมูลเว็บเพจดังกล่าวไปเก็บไว้ในฐานข้อมูลของเสิร์จเอนจิ้น และจะกลับไปตรวจสอบข้อมูลในเว็บไซต์นั้นๆอย่างสม่ำเสมอ เช่น ทุก 1 หรือ 2 เดือนเพื่อสำรวจความเปลี่ยนแปลง

2.1.2 รายการดัชนีข้อมูล (Index หรือ Catalog)

รายการดัชนีข้อมูลที่โปรแกรมรวบรวมเอกสารเว็บพบจะถูกส่งต่อมาจัดเก็บที่รายการดัชนีข้อมูล ซึ่งเปรียบเสมือนเป็นสมุดเล่มใหญ่ที่จัดเก็บสำเนาของเว็บเพจที่โปรแกรมรวบรวมเอกสารเว็บพบ หากข้อมูลที่เว็บไซต์ต้นฉบับมีการเปลี่ยนแปลงข้อมูลในสมุดดัชนีจะเปลี่ยนแปลงด้วย เอกสารเว็บจะถูกทำดัชนีและจัดเก็บตามบัญชีดัชนีที่กำหนดไว้

2.1.3 โปรแกรมการสืบค้น (Search engine software)

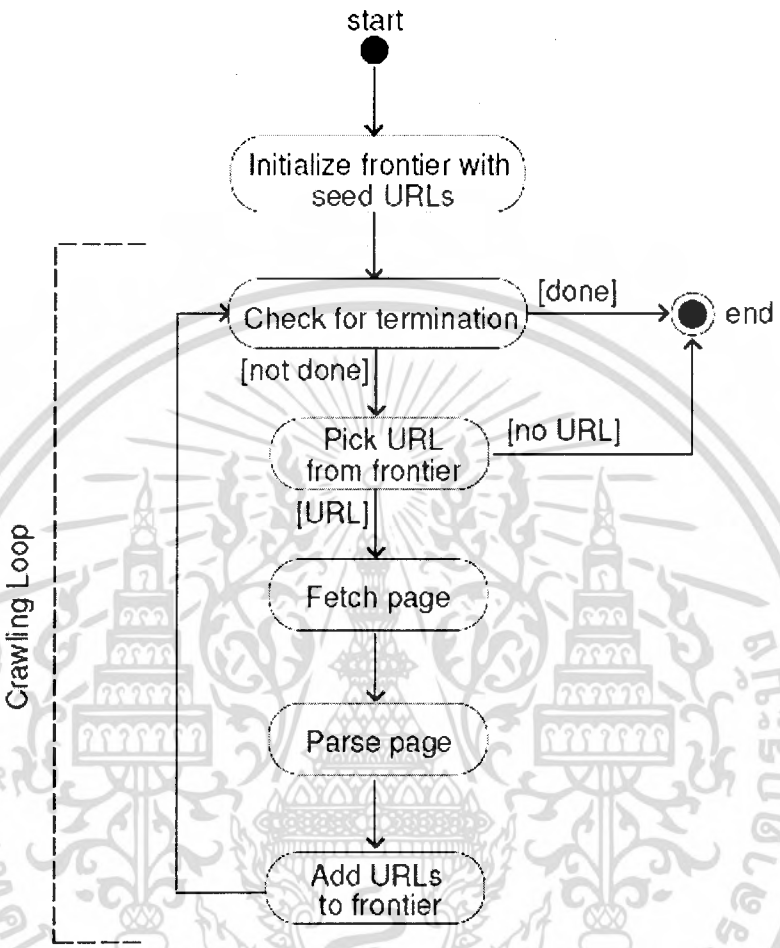
โปรแกรมการสืบค้นจะทำหน้าที่ในการค้นหาข้อมูลจากฐานข้อมูลของเสิร์จเอนจิ้น จะเริ่มต้นการทำงานเมื่อผู้ใช้ป้อนคำค้นหา โปรแกรมจะทำหน้าที่นำคำค้นหาของผู้ใช้ไปจับคู่กับดัชนีในฐานข้อมูลแล้วทำการดึงข้อมูล (เอกสารเว็บ) ที่ตรงกับคำค้นหาออกมา และจัดลำดับผลการค้นหาตามระดับความเกี่ยวข้องที่โปรแกรมประเมินได้ ซึ่งเสิร์จเอนจิ้นแต่ละตัวจะใช้ตรรกะที่แตกต่างกันไป

2.2 หลักการทำงานของโปรแกรมรวบรวมเอกสารเว็บ (Spider หรือ Crawler)

โปรแกรมรวบรวมเอกสารเว็บทำหน้าที่เดินทางไปยังเว็บไซต์ต่างๆเพื่อสะสมเอกสารเว็บ (HTML) เพื่อมาเป็นข้อมูลสำหรับสร้างดัชนีค้นหา โดยทั่วไปแล้วโรบอทจะรวบรวมเอกสารเฉพาะเว็บไซต์ที่ถูกกำหนดอยู่ก่อนแล้ว หลังจากที่โรบอทได้อ่านเว็บเพจเรียบร้อยแล้วโรบอทจะกลับมาอ่านเว็บไซต์ที่อ่านไปแล้ว เพื่อตรวจสอบการเปลี่ยนแปลงตามระยะเวลาที่กำหนด ซึ่งโดยทั่วไปแล้วจะประมาณ 1 หรือ 2 เดือน

โปรแกรมรวบรวมเอกสารเว็บจะทำการเก็บชุดข้อมูลของ URL (Uniform Resource Locator) ที่ยังไม่ได้เข้าถึง เรียกว่า frontier โดยชุดข้อมูลนี้ถูกสร้างขึ้นโดยมี URL เริ่มต้นซึ่งอาจจะได้รับมาจากผู้ใช้หรือโปรแกรมอื่น โดยแต่ละรอบของการค้นหา (crawling loop) จะทำการดึง URL จาก frontier มาทำการค้นหาต่อไปเรื่อยๆ โดยไปดึงหน้าเว็บเพจจาก URL ผ่าน HTTP (Hypertext Transfer Protocol) และทำการวิเคราะห์เนื้อหาในหน้าเว็บเพจที่ได้รับมา เพื่อคัดเอา URL และข้อมูลบางประการออกมาจากเว็บเพจนั้นๆ ถ้าโปรแกรมรวบรวมเอกสารเว็บต้องการจะทำการค้นหาเพจอื่นๆ แต่ใน frontier ว่างเปล่าจะเป็นสัญญาณให้โปรแกรมรวบรวมเอกสารเว็บเข้าสู่สถานะสิ้นสุด (dead-end) ซึ่งหมายถึงโปรแกรมรวบรวมเอกสารเว็บไม่มีหน้าเว็บเพจที่จะทำ

การค้นหาต่อไป ดังนั้นโปรแกรมรวบรวมเอกสารเว็บจะหยุดการทำงาน การทำงานพื้นฐานของโปรแกรมรวบรวมเอกสารเว็บมีกระบวนการทำงานดังรูปที่ 2.1



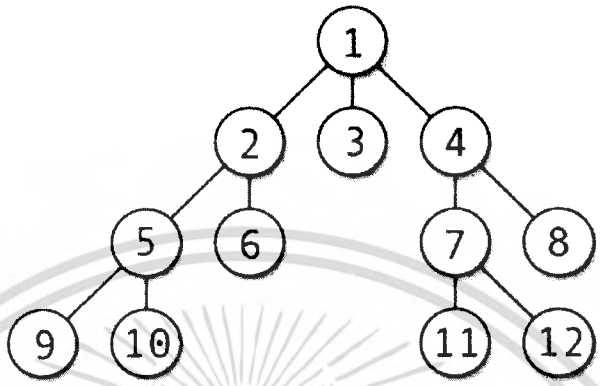
รูปที่ 2.1 การทำงานพื้นฐานของโปรแกรมรวบรวมเอกสารเว็บ

กระบวนการค้นหาสามารถมองเป็นปัญหาการค้นหาแบบกราฟ (graph search problem) ได้ดังรูป 2.2 โดยมองเว็บไซต์เป็นเหมือนกราฟขนาดใหญ่ซึ่งประกอบด้วยเว็บเพจเสมือนโหนด และ hyperlink เป็น edge ของแต่ละโหนด โดยโปรแกรมรวบรวมเอกสารเว็บจะทำการเริ่มที่โหนดจำนวนหนึ่ง (seed) และติดตามไปตามแต่ละ edge เพื่อไปยังโหนดอื่น

กระบวนการในการดึงหน้าเว็บเพจและถอดลิงค์ออกมามีลักษณะที่คล้ายกับการขยายโหนดในการค้นหาแบบกราฟ โดยปกติโปรแกรมรวบรวมเอกสารเว็บจะพยายามไปตาม edge ที่คาดว่าจะพาไปยังส่วนของกราฟที่ตรงกับความต้องการ

Frontier เป็นชุดข้อมูลของโปรแกรมรวบรวมเอกสารเว็บที่บรรจุ URL ของหน้าเว็บเพจที่ยังไม่ได้เข้าถึง บางครั้งอาจจำเป็นต้องเก็บ frontier ลงบนหน่วยความจำ สำหรับโปรแกรมรวบรวมเอกสารเว็บที่มีขอบเขตขนาดใหญ่ โดย frontier อาจจะทำการจัดคิว (queue) โดยกระบวนการ FIFO(first in first out) ซึ่งจะเข้าสู่กรณีการค้นหาแบบ breadth-first search ซึ่ง URL

ที่จะทำการค้นหาต่อไปจะมาจากส่วนหัวของคิว และ URL ใหม่จะถูกเพิ่มลงไปที่ส่วนท้ายของคิว โดยต้องแน่ใจว่าไม่มีการเพิ่ม URL ที่ซ้ำซ้อนลงไปใน frontier เมื่อโปรแกรมรวบรวมเอกสารเว็บพบว่า frontier ว่างเปล่า โปรแกรมรวบรวมเอกสารเว็บจะหยุดการทำงาน ดังแสดงในรูปที่ 2.1



รูปที่ 2.2 การค้นหาแบบ breadth-first และ node ที่จะถูกค้นหาตามลำดับ

ในบางกรณีโปรแกรมรวบรวมเอกสารเว็บอาจจะพบปัญหาสไปเดอร์แทรป (spider trap) ซึ่งหมายถึง URL ที่แตกต่างกันแต่แสดงถึงเว็บเพจเดียวกัน ซึ่งวิธีการหนึ่งที่จะจัดการปัญหานี้คือการกำหนดขอบเขตของการค้นหาไปยังโดเมนที่ได้รับมา

2.3 Google SOAP Search API

บริการของ Google SOAP Search API เป็นบริการที่ให้ผู้พัฒนาซอฟแวร์สามารถค้นหาเว็บเพจได้โดยตรงจากโปรแกรมที่พัฒนาขึ้น ซึ่ง Google ใช้มาตรฐาน SOAP และ WSDL ดังนั้นผู้พัฒนาสามารถพัฒนาโปรแกรมได้จากหลายภาษา เช่น Java, Perl หรือ Visaul Studio .Net

2.3.1 การทำงานของ Google SOAP Search API

2.3.1.1 Search Requests

ในส่วนของ Search Request จะรับคำที่ต้องการค้นหาและกำหนดค่าพารามิเตอร์แล้วส่งค่าพารามิเตอร์นั้นให้กับ Google SOAP Search API และเมื่อ Google ได้รับจะทำการค้นหาและส่งเซ็ตของผลลัพธ์จากการค้นหากลับมา ผลลัพธ์นั้นมาจากอินเด็กซ์ของ Google จากเว็บเพจจำนวนมาก

2.3.1.2 Cache Request

ในส่วนของ Cache Request จะส่ง URL ไปยัง Google SOAP Search API และเมื่อ Google ได้รับจะทำการส่งเนื้อความของ URL นั้นกลับไป เนื้อความนั้นมาจาก Crawler ของ Google ที่เข้าไปตรวจสอบครั้งล่าสุด

2.3.1.3 Spelling Requests

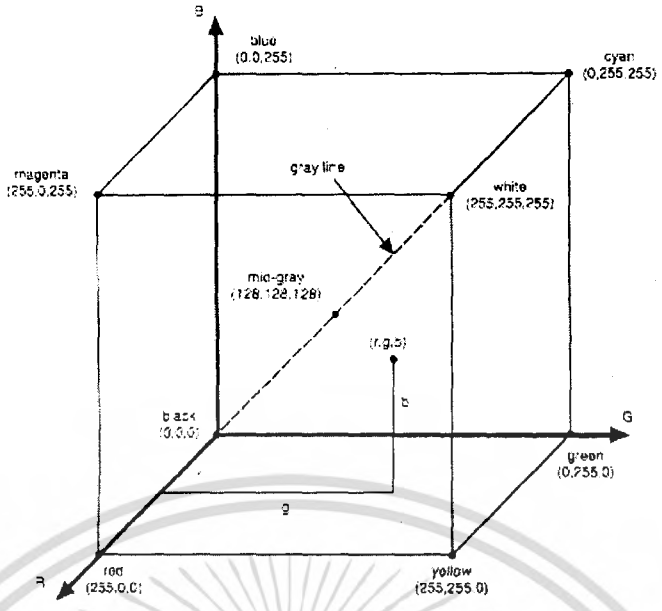
ในส่วนของ Spelling requests จะรับคำที่ต้องการค้นหาเพื่อส่งไปให้ Google SOAP Search API และ Google จะตรวจการสะกดที่ถูกต้องและจะส่งตัวสะกดที่ควรจะเป็นกลับมา (ถ้ามี) การตรวจความถูกต้องจะมีลักษณะเหมือนกับที่พบในหน้าเว็บไซต์ของ Google ค่าที่ส่งกลับมาจะอยู่ในรูปของสตริง

2.4 พื้นฐานและระบบโครงสร้างสี

2.4.1 ระบบโครงสร้างสีอาร์จีบี(RGB Color Model)

ในโครงสร้างนี้ สีแต่ละสีจะปรากฏในรูปแบบของแม่สี คือ สีแดง สีเขียว และ สีน้ำเงิน โครงสร้างสีนี้มีโครงสร้างเป็นลักษณะแกนคาร์ทีเซียนโคออดิเนต (Cartesian coordinate) โดยมีลักษณะเป็นทรงลูกบาศก์ มีค่าสีแดง สีเขียว และสีน้ำเงินอยู่ที่มุมมั้งสามที่เป็นแกน และมีสีฟ้าคราม สีม่วง สีเหลือง อยู่ที่มุมอีก 3 มุม สีดำจะอยู่ที่จุดกำเนิดคือถ้าทุกสีมีค่าเป็น 0 คือเป็น สีดำ สีขาวก็คือตรงข้ามกับสีดำซึ่งอยู่ที่มุนไกลสุดจากจุดกำเนิดในลักษณะทะแยงมุม ในรูปแบบนี้ค่าระดับสีเทา(Gray scale) จะอยู่บนเส้นระหว่างสีดำและสีขาว และสีอื่นๆ ก็จะมีตำแหน่งอยู่ภายในลูกบาศก์นี้ ดังแสดงในรูปที่ 2.3

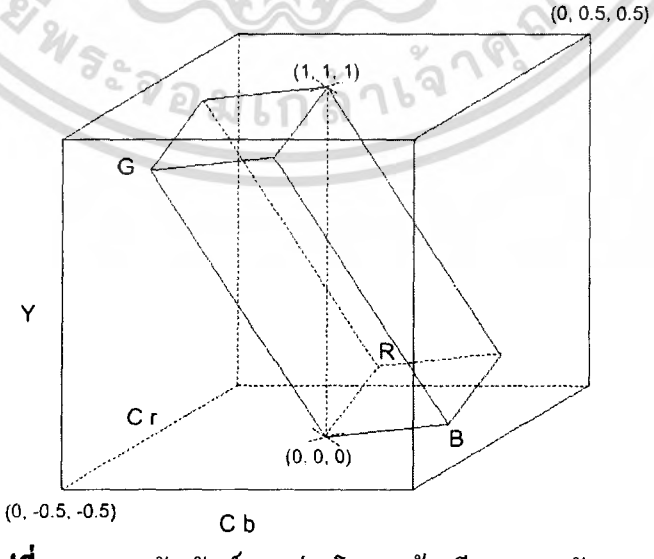
ค่าของสีคือจุดที่อยู่บนพื้นผิวหรือในลูกบาศก์ถูกกำหนดค่าโดยเวกเตอร์ที่ชี้ออกจากจุดกำเนิด ซึ่งช่องว่างแต่ละที่ในลูกบาศก์เรียกว่า ความลึกของพิกเซล (Pixel depth) โดยแม่สีแต่ละสีมีค่า 8 บิต มีได้ 256 ค่า เมื่อรวมสีทั้งหมดของความลึกของพิกเซลในลูกบาศก์(24 บิตอาร์จีบี) จะได้ทั้งหมด $(2^8)^3 = 16,777,216$ สี ภาพในโครงสร้างสีอาร์จีบี ประกอบด้วยภาพสามระนาบที่เป็นอิสระจากกัน



รูปที่ 2.3 โครงสร้างสีอาร์จีบี เป็นลูกบาศก์

2.4.2 ระบบโครงสร้างสีวายซีบีซีอาร์ (YCbCr Color Model)

ในโครงสร้างนี้จะใช้เป็นที่แพร่หลายสำหรับดิจิตอลวีดีโอ ในรูปแบบของโครงสร้างนี้ ค่าปริมาณของแสงในการส่องสว่างจะเก็บข้อมูลนี้ไว้ในส่วนของ (Y) และในส่วนความแตกต่างของสีนั้นจะแบ่งเป็น 2 สีคือ Cb และ Cr โดย Cb จะแสดงให้เห็นถึงความแตกต่างของสีฟ้า และอ้างถึงค่าในหมวดสีฟ้านั้น ส่วน Cr จะแสดงให้เห็นถึงความแตกต่างของสีแดง และอ้างถึงค่าในหมวดสีแดง โดยโครงสร้างสีวายซีบีซีอาร์มีความเที่ยงตรงและแม่นยำมากในส่วนของการส่องสว่าง และหมวดสีซึ่งเป็นโครงสร้างสีที่ใช้กันในการเข้ารหัสแบบเอ็มพีอีจี(MPEG & JPEG) ดังแสดงในรูปที่ 2.4



รูปที่ 2.4 ความสัมพันธ์ระหว่างโครงสร้างสี YCbCr กับ RGB

2.4.3 การแปลงรูปแบบสี

2.4.3.1 การแปลงรูปแบบสีจาก RGB เป็น YCbCr

$$Y = (0.299* R) + (0.587* G) +(0.114* B) \quad (2.1)$$

$$Cb = (- 0.168736 * R)- (0.331264 * G) +(0.5 * B) \quad (2.2)$$

$$Cr = (0.5* R) – (0.418688 * G) – (0.081312 * B) \quad (2.3)$$

โดยกำหนดให้ค่าของ R , G และ B เป็นค่าของ สีแดง สีเขียว และ สีน้ำเงิน ตามลำดับ ซึ่งมีค่าตั้งแต่ 0 ถึง 255 แล้วค่าของ Y จะมีค่าในช่วงตั้งแต่ 0 ถึง 255 และ ค่า Cb ,Cr อยู่ในช่วงตั้งแต่ -128 ถึง 128

2.4.3.2 การแปลงรูปแบบสีจาก YCbCr เป็น RGB

$$R = 1.164*(Y-16) + 1.596*(Cr-128) \quad (2.4)$$

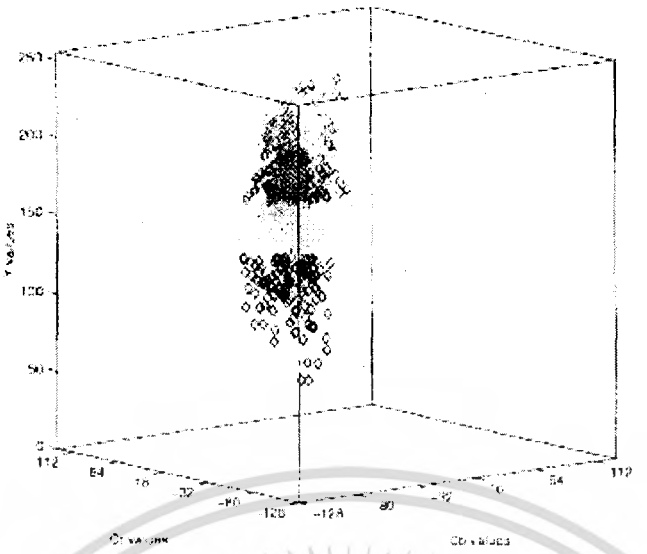
$$G = 1.164*(Y-16)-1.391*(Cb-128)-0.813*(Cr-128) \quad (2.5)$$

$$B = 1.164*(Y-16)+2.018(Cb-128) \quad (2.6)$$

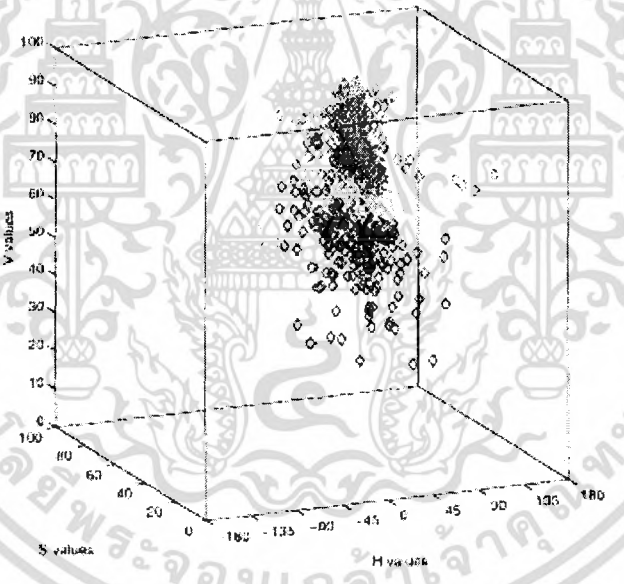
2.5 กระบวนการค้นหาพื้นที่สีผิว

การตรวจหาภาพอนาจารต้องอาศัยกระบวนการค้นหาพื้นที่สีผิว (Detection of Skin color regions) เพื่อแยกสีผิวมนุษย์ออกจากส่วนที่ไม่ใช่สีผิวในรูปภาพ จากงานวิจัยเรื่อง [Face Detection Using Quantized Skin Color Regions Marging and Wavelet Packet Analysis] ในหัวข้อเรื่อง “Skin Color segmentation” เสนอไว้ว่า ในงานวิจัยนั้นใช้โครงสร้างสีสองแบบ คือ YCbCr ซึ่งเป็นโครงสร้างสีที่ใช้กันในการเข้ารหัสแบบ MPEG และ JPEG และใช้โครงสร้างสีแบบ HSV (Hue ,Saturation ,Value)ที่ใช้ในงานคอมพิวเตอร์กราฟฟิก และใกล้เคียงกับการผสมสีของศิลปิน ข้อมูลที่ใช้วิเคราะห์เป็นการสุ่มตัวอย่างสีผิว 950 ตัวอย่าง โดยสุ่มจากหลากหลายเชื้อชาติและมีความสว่างแตกต่างกันไป โดยจะมีการแสดงกลุ่มตัวอย่างสีผิวในแกนของ YCbCr และ HSV ซึ่งแสดงเป็นกราฟข้อมูล 3 มิติ

จากรูปที่ 2.5 และ 2.6 เราสามารถสังเกตเห็นได้ว่า ตัวอย่างสีผิวทั้งใน YCbCr และ HSV ต่างก็มีการกระจายตัวที่มีลักษณะคล้ายกันโดยข้อมูลค่อนข้างที่จะมีการรวมตัวกันอยู่เป็นกลุ่ม ในงานวิจัยนั้นมุ่งที่จะสร้างขอบเขต 3 มิติเพื่อให้คลอบคลุมกลุ่มของตัวอย่างสีผิวที่สุ่มได้มากที่สุดด้วยสมการเส้นตรง



รูปที่ 2.5 กลุ่มตัวอย่างสีผิวในแกนของโครงสร้างสี YCbCr

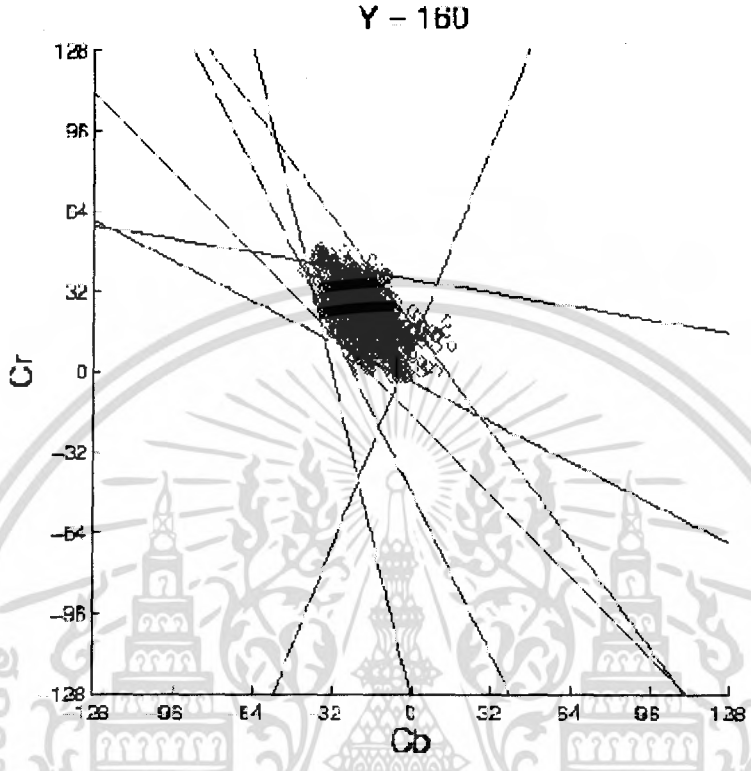


รูปที่ 2.6 กลุ่มตัวอย่างสีผิวในแกนของโครงสร้างสี HSV

ในกราฟ YCbCr พบว่าการเปลี่ยนแปลงของค่า Y มีผลเพียงเล็กน้อยต่อการเปลี่ยนแปลงในระนาบ CbCr โดยทำการพิจารณาใน 2 กรณีคือ ภาพที่มีความมืด (ค่าของ Y มีค่าประมาณ 50) และภาพที่สว่าง (ค่าของ Y มีค่าประมาณ 240)

เมื่อกำหนดค่า Y = 160 เพื่อให้ได้กราฟในระนาบ CbCr และกำหนดสมการเส้นตรงให้ครอบคลุมกลุ่มตัวอย่าง ดังรูปที่ 2.7 การกำหนดค่า Y ค่าเดียวนี้ก่อให้เกิดความผิดผลาดในกระบวนการตรวจหาสีผิว จึงแบ่งช่วง Y ออกเป็นสองช่วงที่จุด Y = 128 ให้มีขอบเขตที่ต่างกัน

โดยกำหนดเป็นตัวแปรที่ต่างกันโดยที่จุด Y = 128 เป็นจุดแบ่งขอบเขตความมืดและสว่างและกำหนดสมการเส้นตรงแปดสมการบนพื้นที่ทั้งสองส่วน



รูปที่ 2.7 กรอบพื้นที่สีผิวโครงสร้างสี YCbCr ที่ระนาบ Y = 160

ทำการคำนวณหาค่า ของ $\theta_1, \theta_2, \theta_3$ และ θ_4 ใน 2 กรณีคือ

เมื่อ $Y > 128$

$$\theta_1 = -2 + \frac{256 - Y}{16}$$

$$\theta_2 = 20 - \frac{256 - Y}{16}$$

$$\theta_3 = 6$$

$$\theta_4 = -8$$

เมื่อ $Y \leq 128$

$$\theta_1 = 6$$

$$\theta_2 = 12$$

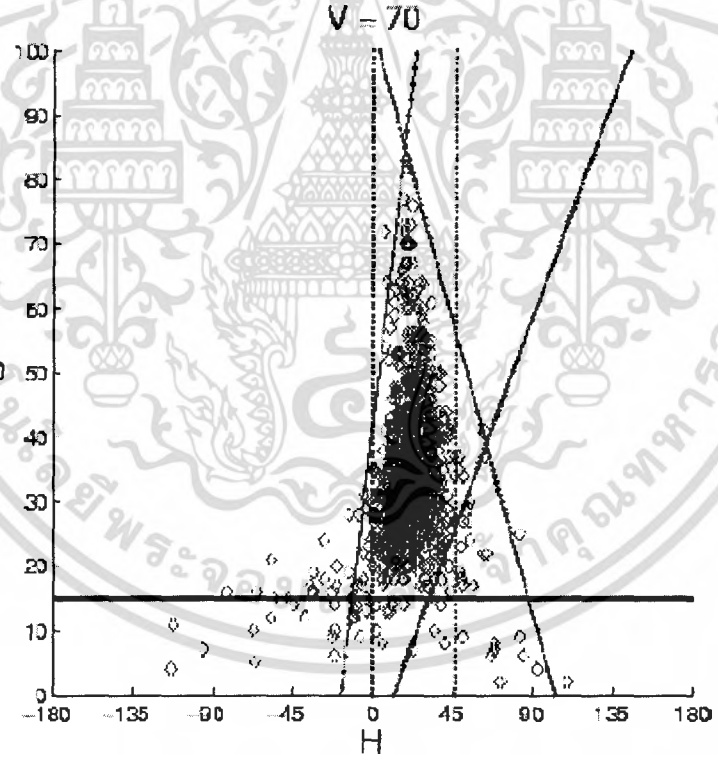
$$\theta_3 = 2 + \frac{Y}{32}$$

$$\theta_4 = -16 + \frac{Y}{16}$$

จากช่วงของค่า Y ทั้งสองช่วง สามารถกำหนดขอบเขตสีผิวได้ด้วยสมการต่อไปนี้

$$Cr \geq -2(Cb + 24)$$
$$Cr \geq -(Cb + 17)$$
$$Cr \geq -4(Cb + 32)$$
$$Cr \geq 2.5(Cb + \theta_1)$$
$$Cr \geq \theta_3$$
$$Cr \geq -0.5(\theta_4 - Cb)$$
$$Cr \leq \frac{220 - Cb}{6}$$
$$Cr \leq \frac{4}{3}(\theta_2 - Cb)$$

สมการเหล่านี้หมายความว่าพิกเซลเป็นสีผิวของมนุษย์ นอกเหนือจากนี้ไม่ใช่สีผิวมนุษย์ โครงสร้างสี YCbCr นั้นเป็นกรรมวิธีที่ดีในการตรวจสอบสีผิวมากกว่า โครงสร้างสี HSV อย่างไรก็ตามก็ยังมีผู้ใช้บางคนใช้การฉายภาพบนระนาบของ HS ซึ่งทำได้โดยการกำหนดค่า thresholds ให้กับ Hue และ Saturation ดังรูปที่ 2.8 ผลของการใช้ค่า thresholds นี้ เราพบว่าผลของการจำแนกสีผิวจะได้รับผลกระทบจากความแตกต่างของค่าความสว่าง



รูปที่ 2.8 กรอบพื้นที่สีผิวโครงสร้างสี HSV ที่ระนาบ V = 70

ในทำนองเดียวกันกับกรณีของ YCbCr ทำการกำหนดค่า V = 70 และได้สมการความสัมพันธ์ของค่า Hue และ Saturation ดังนี้

$$S \geq 10$$
$$V \geq 40$$

$S \leq -H - 0.1V + 110$

ถ้า $H \geq 0$ แล้ว $S \leq 0.08(100-V)H + 0.5V$

ถ้า $H < 0$ แล้ว $S \leq 0.5H + 35$

2.6 กระบวนการปรับความคมชัดของภาพ

จุดประสงค์หลักของการปรับความคมชัดของภาพคือเพื่อทำการเน้นส่วนรายละเอียดในภาพ หรือปรับปรุงรายละเอียดของภาพที่ถูกทำให้พร่ามัว (Blur) อาทิเช่น ความผิดผลาดหรือผลกระทบที่เกิดขึ้นในกระบวนการได้มาซึ่งภาพ (Image acquisition) การใช้กระบวนการของความแตกต่างของสีภายในภาพ (image differentiation) ที่แสดงถึงอัตราความไม่ต่อเนื่องของสีเพื่อตรวจหาเส้นขอบของภาพ และส่วนที่ไม่ต่อเนื่องกันอื่นๆ เช่น สิ่งรบกวน (noise)

กระบวนการปรับความคมชัดของภาพมีพื้นฐานบน first และ second-order derivatives ซึ่งแสดงจุดสนใจบนพื้นที่ที่มีค่าระดับสีเทาคงที่ (*flat segments*) , จุดเริ่มและจุดสิ้นสุดของความไม่ต่อเนื่อง (step and ramp discontinuities) และ พื้นที่ที่มีระดับสีไล่เลี่ยกัน (ramps) โดยมีคุณสมบัติดังนี้

2.6.1 First derivative

Derivativeในฟังก์ชันดิจิตอลนั้นได้ถูกอธิบายในรูปของความแตกต่าง(Difference) โดย First derivative ต้องมีคุณสมบัติดังนี้

- ต้องมีค่าเท่ากับศูนย์บนพื้นที่ที่มีค่าระดับสีเทาคงที่ (flat segments)
- ต้องมีค่าไม่เท่ากับศูนย์ที่จุดเริ่มต้นของพื้นที่ที่มีระดับสีเทาต่างกัน(gray-level step or ramp)
- ต้องมีค่าไม่เท่ากับศูนย์บนพื้นที่ที่มีระดับสีไล่เลี่ยกัน (ramps)

โดย First Order derivative ของฟังก์ชันหนึ่งมิติแนวแกนเอ็กซ์(f(x)) คือ

$$\frac{\partial f}{\partial x} = f(x-1) - f(x) = f'(x) \qquad (2.7)$$

2.6.2 Second derivative

Derivativeในฟังก์ชันดิจิตอลนั้นได้ถูกอธิบายในรูปของความแตกต่าง(Difference) โดย Second derivative ต้องมีคุณสมบัติดังนี้

- ต้องมีค่าเท่ากับศูนย์บนพื้นที่ที่มีค่าระดับสีเทาคงที่ (flat segments)
- ต้องมีค่าไม่เท่ากับศูนย์ที่จุดเริ่มต้นของพื้นที่ที่มีระดับสีเทาต่างกัน(gray-level step or ramp)

- ต้องมีค่าเท่ากับศูนย์บนพื้นที่ที่มีระดับสีไล่เลี่ยกัน (ramps)

โดย Second Order derivative ของฟังก์ชันหนึ่งมิติแนวแกนเอ็กซ์(f(x)) คือ

$$\begin{aligned}\frac{\partial^2 f}{\partial x^2} &= f'(x) - f'(x-1) \\ &= (f(x+1) - f(x)) - (f(x) - f(x-1)) \\ &= f(x-1) + f(x-1) - 2f(x)\end{aligned} \tag{2.8}$$

2.6.3 การใช้ First Derivative ในการเพิ่มคุณภาพของภาพ

First Derivative ในการประมวลผลภาพนั้นเป็นเครื่องมือในการหาค่าความลาดชัน (gradient) สำหรับฟังก์ชัน f(x,y) ค่าความชันของ f ที่ระยะพิกัด (x,y) ถูกแสดงในรูปของเวกเตอร์ที่มีสองคอลัมภ์ดังนี้

$$\nabla f = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{2.9}$$

ในทางปฏิบัติเราสามารถประมาณค่าขนาดความชันได้ด้วยการใช้ค่าสัมบูรณ์ ได้ดังนี้

$$\nabla f \approx |G_x| + |G_y| \tag{2.10}$$

พิจารณารูปที่ 2.9(a) แสดงจุดของภาพที่มีขอบเขต 3x3 พิกเซล (ค่า z เป็นค่าระดับสีเทา) Robert[1965] ได้เสนอคำนิยามในการประมวลผลภาพยุคแรกโดยการใช้ค่าความแตกต่างในแนวทะแยงเป็น $G_x = (z_9 - z_5)$ และ $G_y = (z_8 - z_6)$ และจากสมการที่ 2.10 สามารถคำนวณหาขนาดของความชันได้ดังนี้

$$\nabla f \approx |z_9 - z_5| + |z_8 - z_6| \tag{2.11}$$

สมการที่ 2.11 นี้สามารถใช้ในลักษณะของ masks ในรูปที่ 2.9 (b) และ (c) โดยเรียกว่า Roberts cross-gradient operators

อย่างไรก็ตามการใช้ขนาดของ mask เป็นเลขคู่นั้นทำได้ไม่สะดวก โดยขนาดของ mask ที่เล็กที่สุดที่เราสนใจคือ 3x3 พิกเซลซึ่งเมื่อทำการคำนวณหาขนาดของความชันที่จุดที่มีค่าระดับสีเทาเท่ากับ z_5 แต่ใช้ mask ขนาด 3x3 ได้ดังนี้

$$\begin{aligned}\nabla f \approx\ & |(z7 - 2z8 - z9) - (z1 - 2z2 - z3)| \\ & + |(z3 - 2z6 - z9) - (z1 - 2z4 - z7)|\end{aligned} \tag{2.12}$$

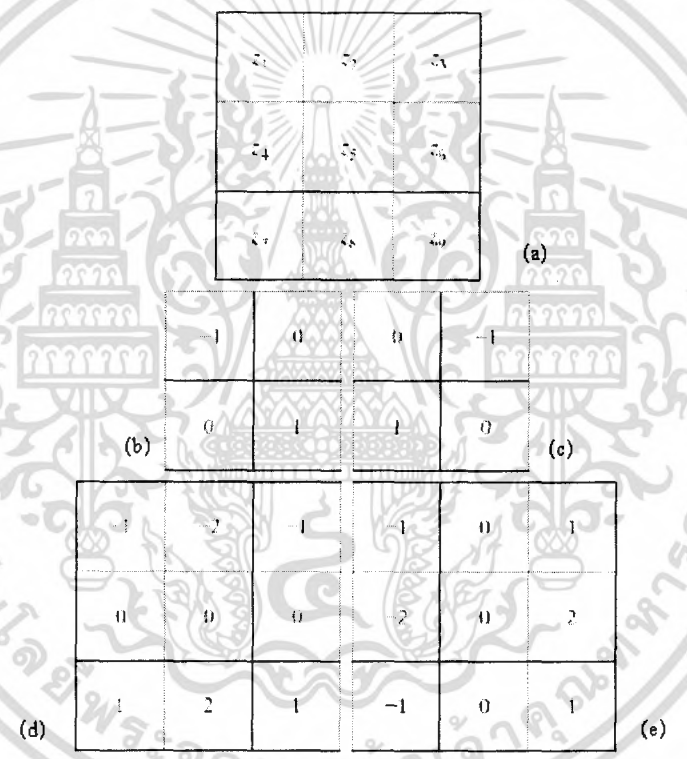
2.6.4 การใช้ Second Derivative ในการเพิ่มคุณภาพของภาพ

ลักษณะของการเพิ่มคุณภาพของภาพด้วย second-order derivatives เราสนใจในการทำตัวกรองแบบไอโซโทรปิก ซึ่งเป็นอิสระต่อทิศทางของความไม่ต่อเนื่องบนภาพ (rotation invariant) กระบวนการแบบไอโซโทรปิกที่ง่ายที่สุดได้แก่ Laplacian ซึ่งมีสมการดังนี้

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{2.13}$$

ซึ่งสามารถแสดงในรูปผลบวกของ 2 องค์ประกอบดังนี้

$$\nabla^2 f = [\, f(x+1, y) - f(x-1, y) + f(x, y-1) + f(x, y-1) \,] - 4f(x, y) \tag{2.14}$$



รูปที่ 2.9 (a) จุดของภาพที่มีขอบเขต 3x3 pixel(ค่า z เป็นค่าระดับสีเทา)

(b)(c) Roberts cross-gradient operators

(d)(e) Sobel operators

2.6.5 การตรวจหาเส้นขอบด้วย กระบวนการโซเบล (Sobel operators)

กระบวนการโซเบล (Sobel Edge Operator) เป็นการตรวจหาขอบของภาพ ซึ่งข้อสันนิษฐานเบื้องต้นผิวของมนุษย์จะมีความราบเรียบต่อเนื่องกันไม่มีขอบเกิดขึ้นภายใน โดยการตรวจหาขอบของ Sobel จะทำโดยการทำ คอนโวลูชั่น (convolution) ซึ่งจะนำพิกเซลที่อยู่ล้อมรอบพิกเซลที่กำลังพิจารณามาทำการคำนวณ โดยเราจะมีตัวเลขคงที่อยู่กลุ่มหนึ่ง เรียกว่า

ค่าตัวร่วม (Mask Coefficient) เข้าไปคูณกับพิกเซลเหล่านั้น จากนั้นนำผลคูณแต่ละตัวมาทำการบวกเข้าด้วยกัน ผลลัพธ์สุดท้ายจะเก็บไว้ในตำแหน่งพิกัดที่กำลังพิจารณาของภาพที่ผ่านการประมวลผลแล้ว

ในทิศทางแนวนอนใช้ Mask Coefficient ขนาด 3x3 ดังรูปที่ 2.9 (d) คือ

$$\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} or \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

และในทิศทางแนวตั้งใช้ Mask Coefficient ขนาด 3x3 ดังรูปที่ 2.9 (e) คือ

$$\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} or \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

อย่างไรก็ตาม Sobel operators สามารถทำได้ด้วยสมการที่ 2.12 เช่นกันโดยการให้ค่าน้ำหนักด้วยสองนั้นเพื่อให้ความสำคัญกับพิกเซลที่กึ่งกลางมากที่สุดโดยค่าน้ำหนักทั้งหมดต้องรวมกันแล้วเท่ากับศูนย์

(a) (b)

รูปที่ 2.10 (a) ตัวอย่างภาพต้นฉบับที่ต้องการหาเส้นขอบ
(b) ภาพที่ถูกตรวจหาเส้นขอบ (Sobel gradient)

2.7 การตรวจสอบเนื้อหาในเว็บไซต์

การแยกแยะระหว่างเว็บไซต์อนาจารกับเว็บไซต์ไม่อนาจารนั้นสามารถทำได้โดยการวิเคราะห์เนื้อหาภายในและโครงสร้างของเอกสาร HTML จากเอกสารอ้างอิง [Identifying and Blocking Pornographic Content] มีการทดลองจากกลุ่มเว็บไซต์ตัวอย่าง สามารถสรุปได้ดังนี้

จากเว็บไซต์อนาจารในกลุ่มตัวอย่างจะประกอบด้วย 3 กลุ่มหลักๆ คือ

1) ในหน้าแรกของเว็บไซต์ มักจะมีข้อความที่สามารถบ่งบอกได้ว่าเว็บไซต์นั้นเป็นเว็บไซต์อนาจาร เช่น เว็บไซต์นี้เหมาะกับบุคคลที่อายุ 18 ปีขึ้นไป เป็นต้น
2) ในหน้าเว็บเพจของเว็บไซต์จะประกอบด้วยภาพอนาจารและคำที่ไม่เหมาะสม
3) ไดเรกทอรีของเว็บไซต์อนาจาร

2.7.1 รูปแบบโครงสร้าง

โครงสร้างเว็บไซต์มีได้มากมายต่างๆกันไม่ว่าจะเป็นเว็บไซต์อนาจารและไม่อนาจาร ซึ่งจะเห็นได้จากตารางที่ 2.1 ค่าบางค่ามีความแตกต่างกันมากซึ่งจะชี้ให้เห็นว่าเว็บไซต์นั้นจัดอยู่ในกลุ่มของเว็บไซต์อนาจาร เมื่อเปรียบเทียบกับกลุ่มเว็บไซต์ที่ไม่อนาจาร ตัวอย่างเช่น มีเพียง 0.8% ของเว็บไซต์ไม่อนาจารเท่านั้นที่มีลิงค์ที่เป็นรูปอยู่มากกว่า 5 ลิงค์ ในขณะที่ 40% ของเว็บไซต์อนาจารจะมีค่าดังกล่าว จากการวิเคราะห์จะแสดงให้เห็นว่า เว็บไซต์อนาจารนั้นจะมีรูปที่เป็นลิงค์เป็นองค์ประกอบนั้นมากกว่า แต่ลิงค์ปกตินั้นจะมีน้อยกว่า อย่างไรก็ตามลักษณะดังกล่าวเป็นเพียงส่วนหนึ่งที่พบในเว็บไซต์อนาจาร ส่วนที่สำคัญส่วนหนึ่งที่ได้ทำการศึกษามานั้นคือระหว่างเว็บไซต์อนาจารกับเว็บไซต์ไม่อนาจารนั้นจะใช้พื้นหลังที่แตกต่างกัน เว็บไซต์ที่ไม่อนาจารนั้นส่วนใหญ่จะใช้สีขาวหรือสีที่สว่างในการทำเป็นพื้นหลัง ส่วนเว็บไซต์ที่อนาจารนั้นจะใช้พื้นหลังที่เป็นสีออกมืด มีขนาดของรูปขนาดเกิน 40*40 พิกเซลที่พบในเว็บไซต์อนาจาร ลิงค์ที่จะไปหน้าเว็บไซต์นั้นจะมีแท็ก หรือ ประกอบอยู่และมีขนาดตัวอักษรที่ใหญ่หรือสีดำ

คุณสมบัติ	**Pornographic**	**Non-pornographic**
จำนวนรูป	74% มีมากกว่า 5 รูป 60% มีมากกว่า 10 รูป	32% มีมากกว่า 5 รูป 13% มีมากกว่า 10 รูป
จำนวนลิงค์ไปยังเว็บไซต์อื่น	73% มีมากกว่า 5 ลิงค์ 46% มีมากกว่า 10 ลิงค์	85% มีมากกว่า 5 ลิงค์ 76% มีมากกว่า 10 ลิงค์
จำนวนลิงค์ภาพและลิงค์ภาพยนตร์	40% มีมากกว่า 5 ลิงค์ 36% มีตั้งแต่ 11 ถึง 20 ลิงค์	0.8% มีมากกว่า 5 ลิงค์ 0.2% มีตั้งแต่ 11 ถึง 20 ลิงค์
จำนวนลิงค์คำ	36% มีน้อยกว่า 40 คำ 28% มีมากกว่า 200 คำ	17% มีน้อยกว่า 40 คำ 50% มีมากกว่า 200 คำ
คำที่เน้น	75% เป็นคำเน้นมากกว่า 10% 21% เป็นคำเน้นทั้งหมด	54% เป็นคำเน้นมากกว่า 10% 9.7% เป็นคำเน้นทั้งหมด

ตารางที่ 2.1 โครงสร้างที่แตกต่างกันระหว่างเว็บไซต์อนาจารและเว็บไซต์ไม่อนาจาร

จากการพิจารณา 2 กลุ่มลักษณะนั้น คำไม่เหมาะสมจะปรากฏอยู่ในส่วนของ Title 98% ของหน้าเว็บไซต์ และอยู่ในส่วนของ Body 97% และมีเพียง 58% ของเว็บไซต์ไม่อนาจารจะใช้ Meta ในการระบุคำสำคัญ หรือรายละเอียดต่างๆ ซึ่งต่างกันเล็กน้อยกับเว็บไซต์อนาจารที่มีเพียง 52% ที่ใช้ Meta ในการระบุคำสำคัญ

ในการวิเคราะห์เนื้อหาของเอกสาร HTML จะต้องพิจารณาลักษณะของแต่ละส่วนเพื่อแยกแยะเว็บไซต์ ซึ่งประกอบด้วย

- Title : จะอยู่ภายใน <TITLE> และ </TITLE> ซึ่งอยู่ในส่วนของ <HEAD>
- Meta : จะอยู่ภายใน <META> และ </META>
- Body : จะเป็นส่วนที่เราเห็นปรากฏอยู่ในเว็บเพจ

จากเอกสารอ้างอิง [Identifying and Blocking Pornographic Content] การทดลองวิเคราะห์คำที่ปรากฏทั้งในเว็บไซต์อนาจารและเว็บไซต์ที่ไม่อนาจาร สามารถสรุปมาเป็นตารางดังตารางที่ 2.2

คำ	จำนวนที่ปรากฏ	
	เว็บไซต์อนาจาร	เว็บไซต์ไม่อนาจาร
Cumshot	136	-
Shemale	108	-
Upskirt	97	-
Gangbang	87	-
Bdsm	63	-
Sex	394	223
Hardcore	392	14
Porn	369	28
Fuck	333	49
babe	288	28
xxx	283	21
pussy	267	1
tit	240	5
girl	493	300
movie	493	488
picture	407	70
video	402	718
gallery	374	391
amateur	355	91
hot	336	474
model	308	405
about	143	3723
article	5	1169

คำ	จำนวนที่ปรากฏ	
	เว็บไซต์อนาจาร	เว็บไซต์ไม่อนาจาร
author	14	809
book	24	1546
busy	31	1411
contact	112	2676
home	131	3254
include	75	1746
industry	17	773
information	100	2913
nation	14	1433
research	4	1082
search	87	2582

ตารางที่ 2.2 คำที่ปรากฏทั้งในเว็บไซต์อนาจารและเว็บที่ไม่ใช่เว็บอนาจาร

2.8 Regular Expression

Regular Expression ถูกนำไปประยุกต์ใช้ในงานด้านต่างๆมากมาย เช่น กลไกการทำงานภายในของคอมไพล์เลอร์ และเว็บไซต์บริการด้านการค้นหาข้อมูลได้นำเอาความสามารถของ Regular Expression เพื่อช่วยในการค้นหาข้อมูล

2.8.1 ความหมายและการใช้งาน

โดยปกติการใช้งาน Regular Expression ใช้เพื่อค้นหาคำที่สอดคล้องกับเงื่อนไข หรืออาจจะแทนที่คำที่สอดคล้องกับเงื่อนไขด้วยข้อความอื่นๆ เป็นต้น ดังนั้นข้อความจะประกอบด้วย 2 ส่วน คือ ส่วนที่เป็นเงื่อนไขหรือที่เรียกว่า Regular Expression และส่วนที่เป็นข้อมูลดิบ

Regular Expression มีความสามารถสูงกว่าการค้นหาโดยทั่วๆไป เนื่องจากการกำหนดเงื่อนไขสามารถใช้สัญลักษณ์พิเศษเพื่อแทนความหมายของข้อความในลักษณะต่างๆได้ สัญลักษณ์เหล่านี้ถูกเรียกว่า “Meta-Character”

2.8.2 Meta-Character ต่างๆภายใน Regular Expression

ไวยกรณ์ต่างๆที่ใช้สร้าง Regular Expression จะใช้อักขระพิเศษที่เรียกว่า Meta-Character ในการแทนเงื่อนไขต่างๆ สัญลักษณ์ทั้งหมดนี้สามารถนำมาเชื่อมต่อเข้าด้วยกัน เพื่อให้ค้นหาคำที่มีความสลับซับซ้อนมากขึ้น

Meta-Character คือ สัญลักษณ์พิเศษที่ถูกใช้ในการกำหนดลักษณะหรือรูปแบบอักษรที่ต้องการค้นหา ได้แก่

1) สัญลักษณ์ (|) เรียกอีกอย่างหนึ่งว่า pipe ทำหน้าที่เป็นทางเลือก มีหลักการทำงานเหมือนปฏิบัติการ OR

2) สัญลักษณ์ (.) ใช้สำหรับตรวจสอบว่าตรงกับอักษรหนึ่งตัว ยกเว้นสัญลักษณ์ขึ้นบรรทัดใหม่

3) สัญลักษณ์ (^) ใช้สำหรับตรวจสอบตำแหน่งเริ่มต้นของสตริงว่าตรงกับคำที่ต้องการหรือไม่

4) สัญลักษณ์ ($) ใช้สำหรับตรวจสอบตำแหน่งสุดท้ายของสตริงว่าตรงกับคำที่ต้องการหรือไม่

5) สัญลักษณ์ (*) ทำหน้าที่กำหนดว่ามีตัวอักษรที่กำหนดหรือไม่ก็ได้

6) สัญลักษณ์ (+) ทำหน้าที่กำหนดว่ามีตัวอักษรที่กำหนดอย่างน้อยหนึ่งตัว

7) สัญลักษณ์ (?) ทำหน้าที่กำหนดว่ามีตัวอักษรที่กำหนดหรือไม่ก็ได้

8) สัญลักษณ์ ({ }) ใช้สำหรับระบุว่าต้องการให้เกิดการซ้ำซ้อนทั้งหมดกี่ตัว

9) สัญลักษณ์ ({ M , N }) ใช้สำหรับระบุว่าต้องการให้เกิดการซ้ำซ้อนทั้งหมดกี่ตัว โดยสามารถกำหนดเป็นช่วงได้ เมื่อ M คือขอบเขตล่างและ N คือขอบเขตบน

10) สัญลักษณ์ ([ ]) ใช้สำหรับกำหนดอักขระที่ต้องการตรวจสอบ โดยให้ตรงกับอักษรที่อยู่ในสัญลักษณ์ [ ]

11) สัญลักษณ์ (-) นำมาใช้ร่วมกับสัญลักษณ์ [] เพื่อสร้างช่วงของตัวอักษรที่จะถูกค้นหา

2.9 การจัดกลุ่มข้อมูล

ในการทดลองจะต้องทำการแบ่งกลุ่มชุดข้อมูลตัวอย่างเพื่อใช้เป็นมาตรฐานในการพิจารณาข้อมูลใหม่ๆว่าเป็นข้อมูลในกลุ่มใด โดยจะทำการแบ่งกลุ่มข้อมูลเป็นสองกลุ่ม คือ กลุ่มของภาพที่ไม่ใช่ภาพอนาจารและกลุ่มของข้อมูลที่เป็นภาพอนาจาร โดยจะใช้หลักการของการวิเคราะห์ส่วนประกอบ, K-Means และ K-Nearest Neighbor (KNN)

หลักการการวิเคราะห์ส่วนประกอบเป็นเทคนิคทางสถิติที่มีประโยชน์ในการทำงานประเภทต่างๆ อาทิเช่น การจดจำใบหน้า(Face recognition) และการบีบอัดภาพ (Image Compression) และเป็นเทคนิคพื้นฐานในการค้นหารูปแบบในข้อมูลที่มีหลายมิติ ก่อนที่จะอธิบายถึงรายละเอียดของหลักการการวิเคราะห์ส่วนประกอบ จะกล่าวถึงทฤษฎีทางคณิตศาสตร์ที่ใช้ในหลักการการวิเคราะห์ส่วนประกอบ ซึ่งประกอบด้วย สถิติ (statistics) ซึ่งพิจารณาการกระจายตัวของข้อมูล และ เมตริกซ์อัลจิบรา (Matrix Algebra) พิจารณาไอแกนเวกเตอร์ (Eigenvectors) และไอแกนแวลูส์(Eigenvalues) ซึ่งเป็นคุณสมบัติสำคัญของเมตริกซ์ที่ซึ่งเป็นรากฐานของหลักการการวิเคราะห์ส่วนประกอบ

2.9.1 สถิติ

สถิติสามารถนำมาใช้เมื่อมีชุดข้อมูลขนาดใหญ่และต้องการวิเคราะห์ความสัมพันธ์ระหว่างแต่ละจุดของชุดข้อมูลนั้น ในที่นี้จะกล่าวถึงการวัดที่กระทำกับกลุ่มของข้อมูล เช่น ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation), วาเรียนซ์ (Variance), โควาเรียนซ์ (Covariance) และโควาเรียนซ์เมตริกซ์ (Covariance Matrix)

- ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation)

ค่าเบี่ยงเบนมาตรฐานของชุดข้อมูลเป็นการวัดการกระจายตัวของข้อมูล คำนิยามของค่าเบี่ยงเบนมาตรฐานคือระยะทางเฉลี่ยจากค่ากลางของชุดข้อมูลถึงจุดข้อมูลหนึ่งๆ การคำนวณค่าเบี่ยงเบนมาตรฐานมีสมการดังนี้

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}} \tag{2.15}$$

การคำนวณหาค่าเบี่ยงเบนมาตรฐานสามารถหาได้จากกำลังสองของระยะทางจากแต่ละจุดของข้อมูลไปยังค่าเฉลี่ยของกลุ่มข้อมูล ทำการรวมค่าที่คำนวณได้ทั้งหมดหารด้วยจำนวนสมาชิกของกลุ่มข้อมูล (n) ลบด้วย 1 แล้วทำการหาค่ารากที่สอง

การที่หารด้วย n-1 ไม่ใช่ n เนื่องจากคำตอบที่ได้มาเป็นการคำนวณหาค่าเบี่ยงเบนมาตรฐานของกลุ่มข้อมูลตัวอย่างแทนการหารด้วยจำนวนข้อมูลทั้งหมด (n)

- วาเรียนซ์ (Variance)

วาเรียนซ์เป็นตัววัดการกระจายตัวของข้อมูลอีกแบบหนึ่ง มีลักษณะแบบเดียวกับค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ซึ่งมีสมการดังนี้

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)} \tag{2.16}$$

ซึ่งอาจกล่าวได้ว่าวาเรียนซ์เป็นกำลังสองของค่าเบี่ยงเบนมาตรฐาน (s^2) โดยทั้งค่าเบี่ยงเบนมาตรฐานและวาเรียนซ์ต่างก็เป็นการวัดการกระจายตัวของข้อมูล แต่ค่าเบี่ยงเบนมาตรฐานจะนิยมใช้มากกว่า

- โควาเรียนซ์ (Covariance)

จากที่กล่าวมาค่าเบี่ยงเบนมาตรฐานและวาเรียนซ์เป็นการวัดการกระจายตัวของข้อมูล 1 มิติ อย่างไรก็ตามมีชุดข้อมูลจำนวนมากที่มีมากกว่า 1 มิติ และจุดประสงค์ของการวิเคราะห์

ข้อมูลทางสถิติเพื่อหาว่าข้อมูลมีความสัมพันธ์ระหว่างมิติหรือไม่ 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

$$\mathrm{var}(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)} \tag{2.17}$$

และสมการของโควาเรียนซ์เป็นดังนี้

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)} \tag{2.18}$$

ถ้าค่าโควาเรียนซ์ที่คำนวณออกมามีค่าเป็นบวกแสดงว่าค่าข้อมูลมีการเพิ่มขึ้นทั้งสองมิติ แต่ถ้าค่าโควาเรียนซ์เป็นลบแสดงว่าข้อมูลค่าเพิ่มขึ้นหนึ่งมิติ ส่วนอีกมิติมีค่าลดลง แต่ถ้าค่าเป็นศูนย์แสดงว่าข้อมูลทั้งสองมิติไม่ได้ขึ้นต่อกัน โดยค่าของ cov(X,Y) จะมีค่าเท่ากับ cov(Y,X) พิจารณาจากสมการ ที่ 2.18

- โควาเรียนซ์เมตริกซ์ (Covariance Matrix)

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

ทางที่ง่ายสำหรับการคำนวณค่าโควาเรียนซ์ระหว่างมิติหลายๆมิติ ทำได้โดยการนำค่าโควาเรียนซ์ทั้งหมดใส่ลงในเมตริกซ์ ซึ่งนิยามสำหรับโควาเรียนซ์เมตริกซ์สำหรับชุดข้อมูลที่มี n มิติ เป็นดังนี้

$$C^{nxn} = (c_{i,j}, c_{i,j} = \mathrm{cov}(Dim_i, Dim_j)) \tag{2.19}$$

ซึ่ง C^{nxn} คือเมตริกซ์ที่มี n แถว n คอลัมภ์ และ Dim_x เป็นมิติที่ x ถ้ามีชุดข้อมูล n มิติ ดังนั้นเมตริกซ์จะมี n แถว n คอลัมภ์ และข้อมูลในเมตริกซ์คือผลลัพธ์ของการคำนวณโควาเรียนซ์

ระหว่างมิติ 2 มิติที่ต่างกัน เช่น ข้อมูลที่อยู่ในแถว 2 คอลัมภ์ 3 คือการคำนวณโควาเรียนซ์ระหว่างมิติที่ 2 และมิติที่ 3

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

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

พิจารณาจากโควาเรียนซ์เมตริกซ์ จะเห็นว่าตามแนวเส้นทะแยงมุมบนซ้ายไปยังล่างขวาเป็นค่าโควาเรียนซ์ระหว่างมิตินั้นและตัวมันเอง หรือก็คือวาเรียนซ์ของมิตินั้น และจาก cov(a,b) มีค่าเท่ากับ cov(b,a) จะเห็นได้ว่าเมตริกซ์นี้ สมมาตรบนเส้นทะแยงมุมนี้ด้วย

2.9.2 เมตริกซ์อัลจิบรา (Matrix Algebra)

ในหัวข้อนี้จะเป็นพื้นฐานของเมตริกซ์อัลจิบราที่ถูกใช้ในหลักการการวิเคราะห์ส่วนประกอบ ซึ่งไอแกนเวกเตอร์และไอแกนแวลูส์ที่จะกล่าวต่อไปมาจากเมตริกซ์อัลจิบรานี้

ตัวอย่างที่ 1 เป็นเมตริกซ์อัลจิบราซึ่งมีนอนไอแกนเวกเตอร์ (non-eigenvector)

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} x \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 11 \\ 5 \end{pmatrix}$$

ตัวอย่างที่ 2 เป็นเมตริกซ์อัลจิบราซึ่งมีไอแกนเวกเตอร์

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} x \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4x \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

- ไอแกนเวกเตอร์ (Eigenvectors)

จากคุณสมบัติของเมตริกซ์ซึ่งสามารถคูณกันได้ระหว่าง 2 เมตริกซ์ ไอแกนเวกเตอร์นั้นเป็นกรณีพิเศษรูปแบบหนึ่ง โดยพิจารณาจากการคูณกันระหว่างเมตริกซ์และเวกเตอร์ในตัวอย่างของหัวข้อเมตริกซ์อัลจิบรา ในตัวอย่างแรกเวกเตอร์ผลลัพธ์ที่ไม่ได้เป็นตัวเลขจำนวนเต็มคูณกับเวกเตอร์ตั้งต้น ต่างกับตัวอย่างที่สองซึ่งผลลัพธ์เป็นสี่เท่าของเวกเตอร์ตั้งต้น จากตัวอย่างจะพบว่าเวกเตอร์ต้นฉบับประกอบด้วย 2 มิติ โดยเวกเตอร์ $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$ จากตัวอย่างที่ 2 แทนเวกเตอร์ที่มีจุดกำเนิดที่ (0,0) ไปยังจุด (3,2)

ไอแกนเวกเตอร์สามารถหาได้จากเมตริกซ์ที่จำนวนคอลัมภ์และจำนวนแถวเท่ากันเท่านั้น แต่ไม่ใช่ทุกเมตริกซ์เช่นนั้นจะมีไอแกนเวกเตอร์ ถ้ากำหนดให้เมตริกซ์ขนาด n x n มีไอแกนเวกเตอร์แล้วจะมี n ไอแกนเวกเตอร์

การหาไอแกนเวกเตอร์ โดยจะทราบว่าความยาวของเวกเตอร์ไม่มีผลต่อการดูว่าเวกเตอร์นั้นเป็นไอแกนเวกเตอร์หรือไม่แต่ทิศทางมีผล ดังนั้นการหาไอแกนเวกเตอร์มักจะมีการทำให้ความยาวของเวกเตอร์เป็นหนึ่ง ซึ่งจะทำให้ทุกไอแกนเวกเตอร์มีความยาวเท่ากัน จากตัวอย่างที่ 2 ไอแกนเวกเตอร์คือ $\begin{pmatrix}3\\2\end{pmatrix}$ และมีความยาวของเวกเตอร์เท่ากับ $\sqrt{(3^2+2^2)}=\sqrt{13}$ เมื่อได้ความยาวของเวกเตอร์แล้วจะนำไปหารเวกเตอร์ตั้งต้น ซึ่งจะทำให้ได้เวกเตอร์ที่มีความยาวเท่ากับหนึ่ง

- ไอแกนแวลูส์(Eigenvalues)

ไอแกนแวลูส์มีความสัมพันธ์ใกล้เคียงกับไอแกนเวกเตอร์จากตัวอย่างที่ 2 ค่าจำนวนเต็ม (4) เรียกว่า ไอแกนแวลูส์ซึ่งสัมพันธ์กับไอแกนเวกเตอร์ ดังนั้นสามารถพบไอแกนแวลูส์และไอแกนเวกเตอร์มาเป็นคู่นั้น หมายถึง เมื่อหาไอแกนแวลูส์ได้ก็จะได้ไอแกนเวกเตอร์ด้วย โดยความสัมพันธ์ของไอแกนแวลูส์ และไอแกนเวกเตอร์ เป็นดังสมการต่อไปนี้

$$Ax = \lambda Ix \tag{2.20}$$

โดย A คือ เมตริกซ์ที่จำนวนคอลัมภ์และจำนวนแถวเท่ากัน

λ คือ เลขจำนวนจริงหรือจำนวนเชิงซ้อน

X คือ เวกเตอร์

I คือ เมตริกซ์เอกลักษณ์ของ A (Identify matric)

จากสมการ 2.20 สำหรับบางเวกเตอร์ X ซึ่ง X ไม่เท่ากับ 0 ค่า λ จะถูกเรียกว่า ไอแกนแวลูส์ของ A และ เวกเตอร์ X จะถูกเรียกว่า ไอแกนเวกเตอร์ของ A และสามารถเขียนสมการ 2.20 ใหม่ได้คือ

$$(A - \lambda I)x = 0 \tag{2.21}$$

จากสมการ 2.21 เราสามารถสรุปจากคุณสมบัติดีเทอมิเนนท์(characteristic determinant)ได้ดังนี้

$$\det(A - \lambda I) = 0 \tag{2.22}$$

จากสมการ 2.22 สามารถคำนวณหา λ ซึ่งเป็น ไอแกนแวลูส์ ของ A ได้ n ค่า (n คือจำนวนมิติของ A) เมื่อได้ λ แทนค่าในสมการ 2.21 จะสามารถคำนวณหา X ซึ่งเป็น ไอแกนเวกเตอร์ของ A ได้ n เวกเตอร์เช่นกัน

2.9.3 หลักการการวิเคราะห์ส่วนประกอบ (Principle Components Analysis)

หลักการการวิเคราะห์ส่วนประกอบเป็นวิธีการระบุถึงรูปแบบของข้อมูลและอธิบายข้อมูล เช่น ความเหมือนหรือแตกต่างของข้อมูล ซึ่งรูปแบบของข้อมูลนั้นหาได้ยากในข้อมูลที่มีมิติจำนวนมากและไม่สามารถแสดงในรูปแบบของกราฟฟิกได้ ซึ่งหลักการการวิเคราะห์ส่วนประกอบเป็นเครื่องมือที่มีประสิทธิภาพมากในการวิเคราะห์ข้อมูล สามารถวิเคระห์รูปแบบของข้อมูลและ ทำให้ข้อมูลเล็กลง ข้อดีของหลักการการวิเคราะห์ส่วนประกอบ คือ เมื่อมีชุด

ข้อมูลที่มีหลายมิติจะสามารถทำการบีบอีดข้อมูลได้โดยการลดมิติลงโดยไม่สูญเสียข้อมูลใดๆ เทคนิคนี้ถูกใช้ในการบีบอัดภาพโดย มีกระบวนการดังนี้

ขั้นตอนที่1 : เก็บรวบรวมข้อมูล

เก็บข้อมูลตามจำนวนมิติที่กำหนด จากรูปที่ 2.3(a) แสดงตัวอย่างข้อมูล 2 มิติ ได้แก่ x และ y ของข้อมูล 10 ตัวอย่าง

DATA		DATA ADJUST	
X	Y	X	Y
2.5	2.4	0.69	0.49
0.5	0.7	-1.31	-1.21
2.2	2.9	0.39	0.99
1.9	2.2	0.09	0.29
3.1	3.0	1.29	1.09
2.3	2.7	0.49	0.79
2	1.6	0.19	-0.31
1	1.1	-0.81	-0.81
1.5	1.6	-0.31	-0.31
1.1	0.9	-0.71	-1.01

(a) (b)

ตารางที่ 2.3 (a) ตัวอย่างข้อมูล 2 มิติ (x,y) ของข้อมูล 10 ตัวอย่าง
(b)) ตัวอย่างข้อมูล2 มิติ ที่มีการปรับค่าแล้วของข้อมูล 10 ตัวอย่าง

ขั้นตอนที่ 2 : ลบจากค่ากลางของชุดข้อมูลนั้นๆ

การนำข้อมูลลบจากค่ากลางหมายถึงการเฉลี่ยค่าข้อมูลของแต่ละมิติ โดยขั้นแรกต้องทำการหาค่ากลางของชุดข้อมูลแต่ละมิติ และทำการลบข้อมูลทุกๆค่าในมิตินั้นๆด้วยค่ากลางของมิตินั้นๆ จะทำให้ได้ชุดข้อมูลซึ่งมีค่ากลางเท่ากับ 0 ค่าของข้อมูลที่ทำการลบออกจากค่ากลางแล้วจะเรียกว่า ข้อมูลที่มีการปรับค่าแล้ว (Data Adjust) ดังรูปที่ 2.3(b) แสดงตัวอย่างข้อมูล2 มิติ ที่มีการปรับค่าแล้วของข้อมูล 10 ตัวอย่างในรูปที่ 2.3(a)

ขั้นตอนที่ 3 : คำนวณค่าโควาเรียนซ์เมตริกซ์

ทำการคำนวณหาโควาเรียนซ์เมตริกซ์ของ ข้อมูลที่ปรับค่าแล้วตามที่ได้กล่าวไว้ข้างต้น ซึ่งจะได้โควาเรียนซ์เมตริกซ์ขนาด n x n (n คือจำนวนมิติทั้งหมด) ตัวอย่างจากตารางที่ 2.3(b) ทำ

การคำนวณหาโควาเรียนซ์เมตริกซ์ของตัวอย่างข้อมูล2 มิติ ที่มีการปรับค่าแล้วจากสมการที่ 2.9 ซึ่งจะได้โควาเรียนซ์เมตริกซ์ขนาด 2x2 ดังนี้

$$\text{cov} = \begin{pmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555556 \end{pmatrix}$$

ขั้นตอนที่ 4 : คำนวณหาไอแกนเวกเตอร์และไอแกนแวลูส์ของโควาเรียนซ์เมตริกซ์

เนื่องจากโควาเรียนซ์เมตริกซ์เป็นเมตริกซ์ที่มีจำนวนแถวและจำนวนคอลัมภ์เท่ากัน ดังนั้นสามารถคำนวณหาไอแกนเวกเตอร์และไอแกนแวลูส์สำหรับเมตริกซ์นั้นได้ โดยไอแกนเวกเตอร์เป็นเวกเตอร์ 1 หน่วย จากตัวอย่างในขั้นตอนที่ 3 นำโควาเรียนซ์เมตริกซ์ขนาด 2x2 มาคำนวณหาไอแกนเวกเตอร์และไอแกนแวลูส์ จากสมการ 2.21 และ 2.22 ได้ดังนี้

$$eigenvalues = \begin{pmatrix} 0.0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -0.735178656 & -0.677873399 \\ 0.677873399 & -0.735178656 \end{pmatrix}$$

ขั้นตอนที่ 5 : เลือกองค์ประกอบและแปลงให้อยู่ในรูปเมตริกซ์ของเวกเตอร์

ขั้นตอนในการบีบอัดข้อมูลและทำการลดมิติจากการพิจารณาไอแกนเวกเตอร์และไอแกนแวลูส์ เลือกค่าไอแกนเวกเตอร์ที่มีค่าไอแกนแวลูส์สูงที่สุดซึ่งแสดงถึงความสัมพันธ์ของข้อมูลซึ่งมีความสำคัญ

หากทำการเรียงข้อมูลตามค่าของไอแกนแวลูส์จากมากไปหาน้อยจะได้องค์ประกอบที่เรียงลำดับความสำคัญ ดังนั้นทำให้สามารถตัดสินใจที่จะละเลยองค์ประกอบที่มีความสำคัญน้อยกว่า ซึ่งจะทำให้สูญเสียข้อมูลบางส่วน แต่ถ้าจำนวนของไอแกนแวลูส์น้อยจะทำให้สูญเสียข้อมูลไม่มาก ถ้าตัดองค์ประกอบบางส่วน ชุดของข้อมูลสุดท้ายจะมีมิติน้อยกว่าต้นฉบับ ในทางปฏิบัติถ้าต้นฉบับมี n มิติ เราจะคำนวณได้ n ไอแกนเวกเตอร์และ n ไอแกนแวลูส์และเลือกมาเฉพาะ p ไอแกนเวกเตอร์แรก และชุดข้อมูลสุดท้ายจะมีเพียง p มิติ จากนั้นต้องทำการแปลงให้อยู่ในรูปของเมตริกซ์ของเวกเตอร์โดยการนำค่าไอแกนเวกเตอร์ที่เลือกมาแปลงเป็นเมตริกซ์ให้ไอแกนเวกเตอร์อยู่ในคอลัมภ์ จากตัวอย่างไอแกนเวกเตอร์และไอแกนแวลูส์ที่คำนวณได้ทำการเรียงตามค่าของไอแกนแวลูส์จากมากไปหาน้อยจะได้ค่าดังนี้

$$\begin{pmatrix} -0.677873399 & -0.735178656 \\ -0.735178656 & 0.677873399 \end{pmatrix}$$

ซึ่งสามารถพิจารณาตัดเวกเตอร์ที่มีไอแกนแวลูส์ต่ำกว่าออกได้ดังนี้

$$\begin{pmatrix} -0.677873399 \\ -0.735178656 \end{pmatrix}$$

ขั้นตอนที่ 6 : ได้ข้อมูลชุดใหม่

เป็นขั้นตอนสุดท้ายของหลักการการวิเคราะห์ส่วนประกอบจากไอแกนเวกเตอร์ที่เลือกไว้ จากนั้นทำการทรานสโพสเวกเตอร์ (transpose vector) และนำมาคูณทางซ้ายของข้อมูลต้นฉบับ ซึ่งมีสมการดังนี้

$$\text{FinalData} = \text{RowFeatureVector} \times \text{RowDataAdjust} \qquad (2.23)$$

ซึ่ง RowFeatureVector คือเมตริกซ์ที่มีไอแกนเวกเตอร์อยู่ในคอลัมภ์ที่ถูกทรานสโพส ดังนั้นไอแกนเวกเตอร์จะอยู่ในแถวของเมตริกซ์แทน ไอแกนเวกเตอร์ที่ถูกทรานสโพสและมีค่าความสำคัญสูงสุดจะอยู่ส่วนบนสุดของ RowFeatureVector และ RowDataAdjust เป็นชุดข้อมูลที่ถูกลบด้วยค่ากลางและถูกทรานสโพส และ FinalData เป็นชุดข้อมูลสุดท้ายที่ประกอบด้วยไอเท็มของข้อมูล (data item)ในคอลัมภ์ และมิติในแต่ละแถว จากตัวอย่างสามารถคำนวณหา FinalData จากสมการ 2.23 ได้ดังนี้

$$\text{FinalData} = \begin{pmatrix} -0.677873399 & -0.735178656 \end{pmatrix} \times \begin{pmatrix} 0.69 & -1.31 & .... & -0.31 & -0.71 \\ 0.49 & -1.21 & .... & -0.31 & -1.01 \end{pmatrix}$$

เมื่อทำการคำนวณจะได้ FinalData ซึ่งเหลือเพียง 1 มิติดังตารางที่ 2.4

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

PCA

FinalData
-0.827970186
1.77758033
-0.274210416
-0.992197494
-1.67580142
-0.912949103
0.991094375
1.14457216
0.438046137
1.22382056

ตารางที่ 2.4 ตัวอย่างข้อมูลชุดใหม่ที่ผ่านกระบวนการวิเคราะห์ส่วนประกอบ

2.9.4 K-Means

K-Means เป็นอัลกอริทึมที่ใช้แยกข้อมูลออกเป็นกลุ่มๆโดยอาศัยคุณลักษณะต่างๆเพื่อแยกข้อมูลออกเป็น K กลุ่ม การจัดกลุ่มแบบ K-Means ทำได้โดย

1. หาจุดเซ็นทรอยด์ (Centroid)
2. หาระยะทางของแต่ละข้อมูลกับจุดเซ็นทรอยด์
3. จัดกลุ่มข้อมูลไปยังกลุ่มที่มีระยะทางจากจุดเซ็นทรอยด์ของกลุ่มนั้นสั้นที่สุด

ในตอนแรกต้องกำหนดจุดเซ็นทรอยด์ตั้งต้นไว้ก่อน โดยการเลือกจุดข้อมูลใดๆในชุดข้อมูลนั้น ในการทดลองจะทำการกำหนดจุดเซ็นทรอยด์ 3 จุด จากนั้นทำการคำนวณระยะทางจากข้อมูลแต่ละตัวถึงจุดเซ็นทรอยด์ทั้งสามจุด โดยการใช้การคำนวณระยะทางแบบยูคลิดีน (Euclidean distance) โดยมีสมการดังนี้

$$Dis\tan ce = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2} \tag{2.24}$$

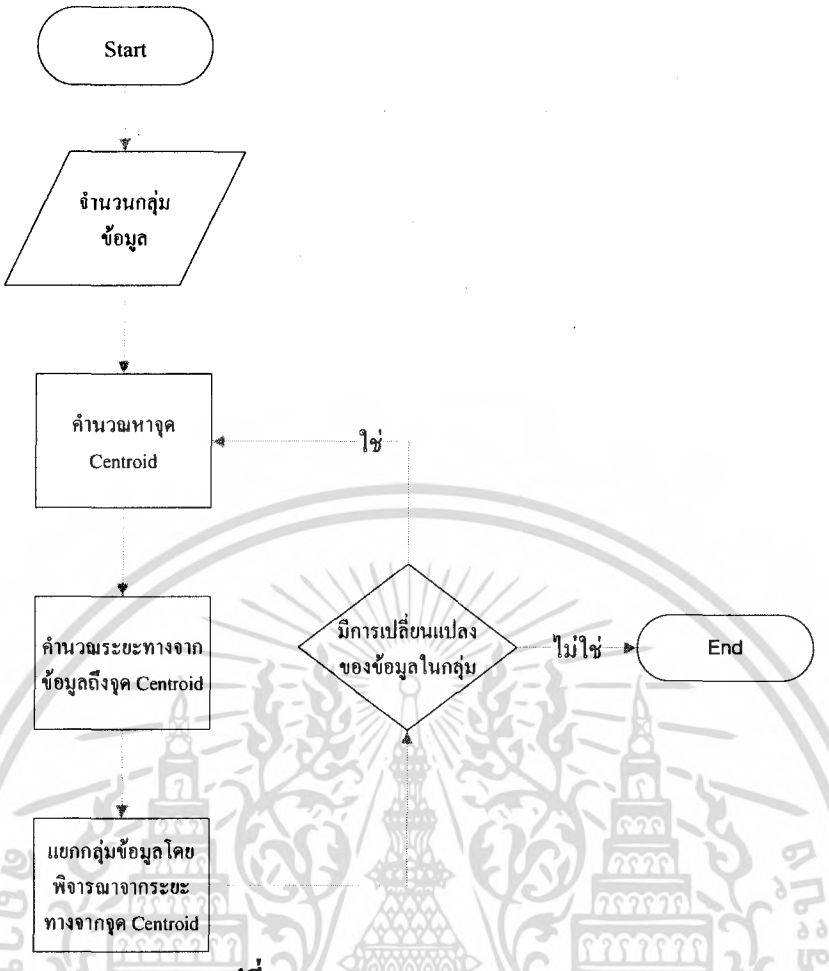
เมื่อคำนวณหาระยะทางจากข้อมูลแต่ละตัวกับจุดเซ็นทรอยด์ทั้งสามจุดแล้ว จะทำการแยกข้อมูลแต่ละตัวไปยังกลุ่มที่มีระยะทางจากจุดเซ็นทรอยด์ของกลุ่มนั้นน้อยที่สุด เมื่อได้ชุดข้อมูลของกลุ่มที่1 ,2 ,3 มาแล้ว ก็จะต้องทำการคำนวณหาจุดเซ็นทรอยด์ของกลุ่มนั้นใหม่โดยดูจากชุดข้อมูลในกลุ่มนั้นๆ คำนวณหาจุดเซ็นทรอยด์ โดยการหาค่าเฉลี่ยของค่า x และ y ของข้อมูลแต่ละตัวในกลุ่มนั้น

$$c = \left(\frac{x_1 + x_2 + ... + x_n}{m}\right), \left(\frac{y_1 + y_2 + ... + y_n}{m}\right) \tag{2.25}$$

n = จำนวนข้อมูลของกลุ่มนั้น

m = จำนวนจุด เซ็นทรอยด์

เมื่อได้จุดเซ็นทรอยด์ใหม่ของแต่ละกลุ่มแล้วก็ทำการคำนวณหาระยะทางแบบเดิมทำวนไปเรื่อยๆจนกว่าค่าของจุดเซ็นทรอยด์จะคงที่ การหา K-Means แสดงขั้นตอนการทำได้ดังรูปที่ 2.11



รูปที่ 2.11 flow chart ของการหา K-Means

2.9.5 K-Nearest Neighbor (KNN)

K-Nearest Neighbor เป็นการเรียนรู้แบบหนึ่ง ซึ่งใช้ประกอบกับโปรแกรมประเภทเหมืองข้อมูล, สถิติของการจดจำรูปแบบ, การประมวลผลภาพ เป็นต้น อัลกอริทึมนี้มีจุดประสงค์เพื่อจัดกลุ่มข้อมูลโดยพิจารณาจากลักษณะเฉพาะและกลุ่มตัวอย่าง วิธีการจัดกลุ่มไม่ได้ใช้กลุ่มตัวอย่างในการกำหนดผลทดลอง แต่การจัดกลุ่มจะใช้คะแนนส่วนใหญ่ของการจัดกลุ่ม K อ็อบเจ็กต์ อัลกอริทึม K-Nearest Neighbor เป็นการจัดกลุ่มจากการคาดการณ์ค่าของจุดคงที่ วิธีการคำนวณอัลกอริทึมของ KNN มีขั้นตอนดังนี้

1. พิจารณาพารามิเตอร์ K = จำนวนของ nearest neighbor
2. คำนวณระยะทางระหว่างจุดเซ็นทรอยด์ที่หามาจาก K-means กับข้อมูล
3. เรียงลำดับระยะทางและพิจารณา nearest neighbor โดยมีพื้นฐานบน K-th minimum distance
4. รวมกลุ่มของ nearest neighbor เข้าด้วยกัน
5. ใช้กลุ่มหลักของ nearest neighbor เป็นค่าที่ไว้คาดการณ์ความน่าจะเป็นของข้อมูลใหม่ๆว่าเป็นกลุ่มไหน

บทที่ 3
การออกแบบและพัฒนา

โครงงานระบบตรวจหาข้อมูลอันตรายในเวิร์ลไวด์เว็บ ใช้กระบวนการตรวจหาเว็บไซต์ด้วยกลไกการทำงานของโปรแกรมรวบรวมเอกสารเว็บ ที่มุ่งเน้นในด้านประสิทธิภาพความรวดเร็วของการตรวจหาและการรวบรวมลิงค์ให้ได้จำนวนมากและครอบคลุมทั่วถึง เมื่อได้ลิงค์ของเว็บเพจ ระบบจะทำการวิเคราะห์เนื้อหาภายในเว็บเพจทั้ง รูปภาพ และข้อความสำคัญ เพื่อระบุว่าเว็บเพจนั้นๆ มีระดับความไม่เหมาะสมเท่าใด ให้มีความถูกต้องและรวดเร็ว และทำการจัดเก็บรายการเว็บไซต์ที่ตรวจพบว่ามีเนื้อหาไม่เหมาะสมไว้ในรายการเว็บไซต์ต้องห้ามและทำการตรวจสอบเว็บเพจนั้นๆ อย่างสม่ำเสมอ เพื่อให้ข้อมูลมีความทันสมัยอยู่ตลอดเวลา

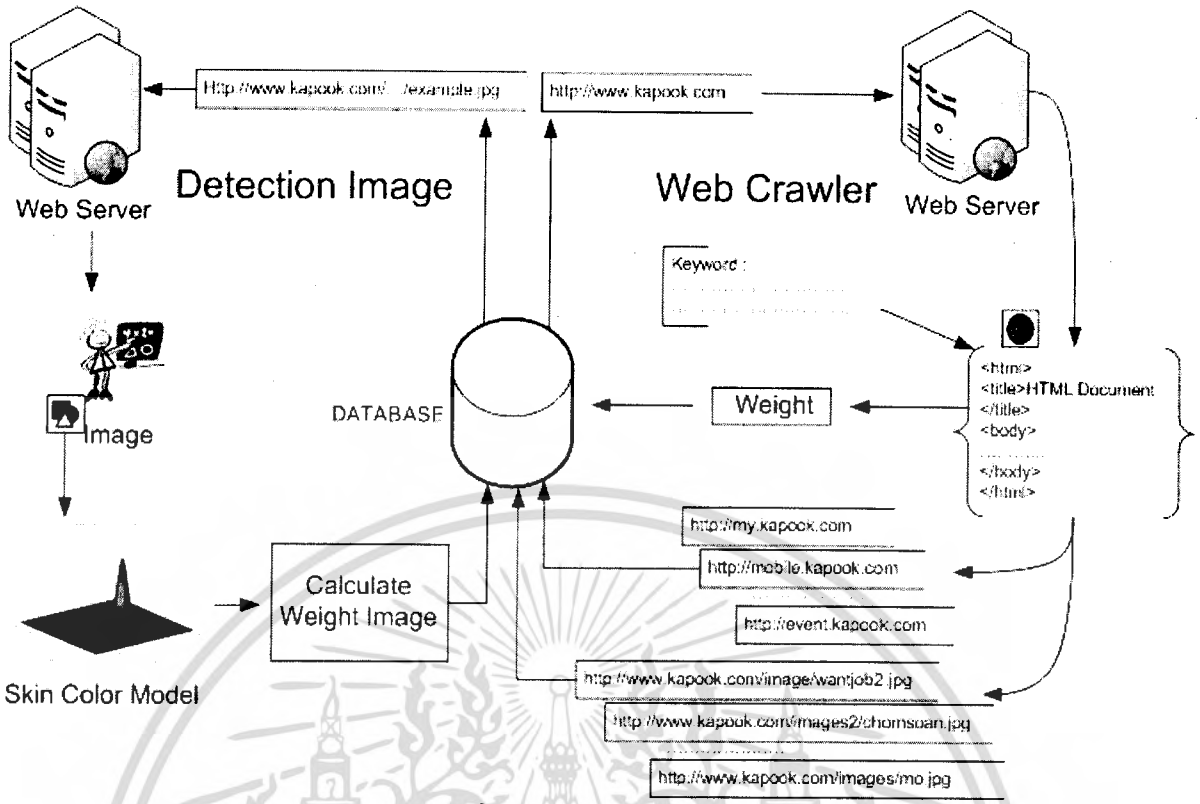
โครงงานนี้ได้เลือกใช้ภาษา ไพธอน (Python) ในการเขียนโปรแกรมเพื่อให้สามารถพัฒนาโปรแกรมให้มีขีดความสามารถที่สูง และสามารถใช้งานได้หลากหลาย ทั้งส่วนของโปรแกรมรวบรวมเอกสารเว็บ, การวิเคราะห์เนื้อหาในเว็บเพจทั้งรูปภาพ และข้อความสำคัญ รวมทั้งส่วนที่ทำการติดต่อกับฐานข้อมูล

3.1 โครงสร้างของโครงงาน

ระบบนี้ได้แบ่งเป็น 3 ส่วนหลักคือ

1) โปรแกรมรวบรวมเอกสารเว็บ (Spider หรือ Crawler)
2) โปรแกรมตรวจสอบภาพอนาจาร (Pornographic detection)
3) โปรแกรมตรวจสอบเนื้อหาของเว็บไซต์ (Text detection)

จากรูปที่ 3.1 แสดงการทำงานของระบบ โดยเริ่มต้นการทำงานด้วย โปรแกรมรวบรวมเอกสารเว็บเพื่อทำการค้นหาลิงค์ของเว็บเพจอื่น และลิงค์ของภาพที่ปรากฏในเว็บเพจนั้น เมื่อโปรแกรมทำการดึงข้อมูลเหล่านี้ออกมาได้ จะทำการเพิ่มข้อมูลเข้าสู่ฐานข้อมูลพร้อมทั้งทำการปรับค่าเวลาของการเข้าถึงเว็บเพจนั้นๆใหม่ ในระหว่างการดึงข้อมูลเกี่ยวกับลิงค์ โปรแกรมจะทำการวิเคราะห์เนื้อหาของเว็บเพจนั้นพร้อมคำนวณค่าความไม่เหมาะสมด้วยคำสำคัญเพื่อประกอบการพิจารณาในขั้นตอนต่อไป



รูปที่ 3.1 การทำงานของระบบ

จากนั้นโปรแกรมวิเคราะห์ข้อมูลภาพจะทำการดึงลิงค์ของภาพจากฐานข้อมูล และทำการเข้าไปยัง เครื่องที่ให้บริการ เพื่อร้องขอข้อมูลภาพเมื่อได้ภาพที่ต้องการ โปรแกรมจะทำการวิเคราะห์ค่าของสีภายในภาพ และดึงข้อมูลเพื่อใช้ในการพิจารณาภาพ และบันทึกลงฐานข้อมูลว่ารูปนั้นเป็นรูปอนาจารหรือไม่

ส่วนสุดท้ายโปรแกรมจะทำการคำนวณหาค่าน้ำหนักของความไม่เหมาะสมของแต่ละเว็บเพจ และเพิ่มรายชื่อของเว็บไซต์ที่มีค่าน้ำหนักของความไม่เหมาะสมเกินกว่าเกณฑ์ที่กำหนดลงในรายการเว็บไซต์ต้องห้ามในฐานข้อมูลเพื่อให้ผู้ใช้สามารถเรียกดูข้อมูลดังกล่าวได้โดยโปรแกรม ทั้งสองจะทำงานจนกว่าจะมีการสั่งหยุดการทำงาน

3.2 โปรแกรมรวบรวมเอกสารเว็บ (Spider หรือ Crawler)

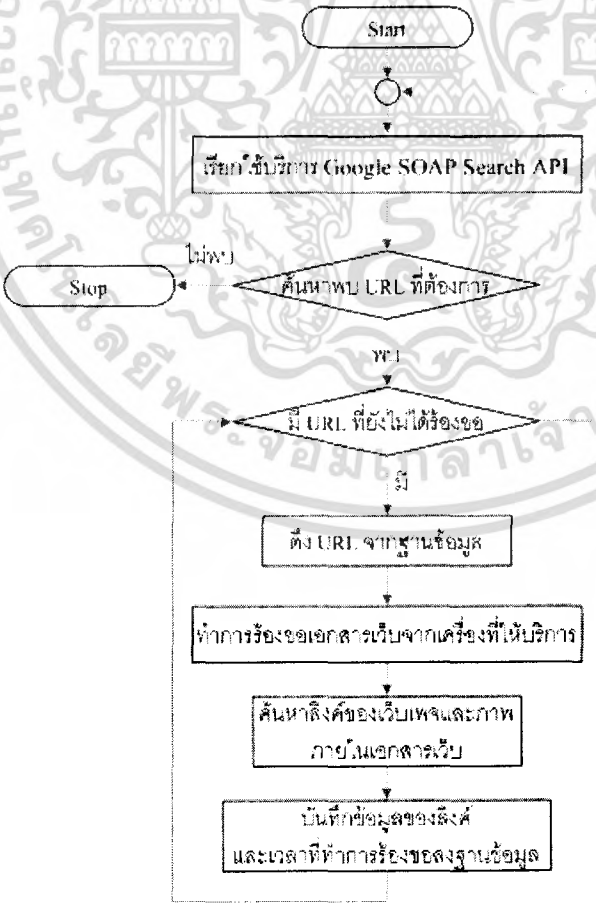
โปรแกรมรวบรวมเอกสารเว็บทำหน้าที่รวบรวมลิงค์ให้ได้มากและครอบคลุมที่สุดเพื่อให้เกิดการกระจายการค้นหาที่มีประสิทธิภาพถือว่าเป็นส่วนที่สำคัญที่สุดของระบบเสิร์จเอนจิ้นที่เน้นเรื่องประสิทธิภาพและความเร็วของการค้นหา โปรแกรมรวบรวมเอกสารเว็บต้องมีการติดต่อกับเครือข่ายภายนอกเพื่อทำการร้องขอเอกสารเว็บและนำมาวิเคราะห์ โครงงานนี้ได้จัดทำโปรแกรมรวบรวมเอกสารเว็บที่ทำการค้นหาลิงค์แบบต่อเนื่องและบันทึกข้อมูลเกี่ยวกับเอกสารเว็บลงฐานข้อมูล

โปรแกรมรวบรวมเอกสารเว็บที่มีการติดต่อกับฐานข้อมูลทั้งการเรียกดูข้อมูล, การบันทึกข้อมูล และการปรับค่าข้อมูลให้ทันสมัย จึงจำเป็นอย่างยิ่งที่โปรแกรมรวบรวมเอกสารเว็บต้องมีส่วนที่ทำการติดต่อกับฐานข้อมูลที่มีประสิทธิภาพ

โปรแกรมรวบรวมเอกสารเว็บนั้นมีระบบการทำงานดังนี้

- เรียกดูรายชื่อ URL ของเว็บเพจจากฐานข้อมูล
- ติอต่อไปยังเครื่องที่ให้บริการเว็บเพื่อร้องขอเอกสารเว็บตามรายชื่อ URL
- ทำการค้นหาลิงค์ของเว็บเพจและรูปภาพภายในเอกสารเว็บที่ได้รับมา
- วิเคราะห์เนื้อหาของเอกสารเว็บ ด้วยคำสำคัญ
- บันทึกข้อมูลของลิงค์และเวลาที่ทำการร้องขอลงฐานข้อมูล

ส่วนประกอบสำคัญของโปรแกรมรวบรวมเอกสารเว็บ คือ รายการเว็บไซต์ที่ยังไม่ได้เข้าถึง (frontier) ที่ทำหน้าที่เก็บรายชื่อ URL ที่ยังไม่ได้เข้าถึง การเรียกดูรายชื่อ URL จาก frontier จำเป็นอย่างยิ่งที่ต้องมีการกำหนด URL ตั้งต้นในฐานข้อมูลเพื่อให้โปรแกรมรวบรวมเอกสารเว็บสามารถเริ่มต้นการทำงานได้ โดยออกแบบให้เริ่มต้นการทำงานโดยร้องขอบริการจาก Google SOAP Search API โดยใช้คำสำคัญในการค้นหาเพื่อขอรายชื่อเว็บไซต์ที่เข้าข่ายและ กำหนดให้เป็น URL ตั้งต้นในรายการเว็บไซต์ที่ยังไม่ได้เข้าถึง (frontier)



รูปที่ 3.2 ขั้นตอนการทำงานของโปรแกรมรวบรวมเอกสารเว็บ

จากรูปที่ 3.2 แสดงขั้นตอนการทำงานของโปรแกรมรวบรวมเอกสารเว็บ ในแต่ละ รอบการค้นหา (crawling loop) จะทำการดึง URL ที่ต้องการค้นหาต่อไปจาก รายการเว็บไซต์ที่ยังไม่ได้เข้าถึง (frontier) และทำการร้องขอเอกสารเว็บของ URL นั้นๆ ผ่าน โปรโตคอล HTTP และทำการแบ่งแยกเอกสารเว็บเพื่อดึง URLหรือข้อมูลอื่นที่ต้องการและสุดท้าย ทำการเพิ่ม URL ที่ยังไม่ได้เข้าถึงเข้าไปยัง frontier

3.2.1 รายการเว็บไซต์ที่ยังไม่ได้เข้าถึง (frontier)

ออกแบบให้รายการเว็บไซต์ที่ยังไม่ได้เข้าถึง (frontier) มีลักษณะเป็นคิวแบบเข้าก่อนออกก่อน (FIFO queue) ในลักษณะการค้นหาแบบแนวกว้าง (breadth-first search) โดย URL ที่จะถูกพิจารณาจะมาจากส่วนหัวของคิว และ URL ใหม่จะถูกเพิ่มที่ส่วนท้ายของคิว และต้องแน่ใจว่าไม่มีการเพิ่ม URL ที่ซ้ำกันบนรายการเว็บไซต์ที่ยังไม่ได้เข้าถึง

เมื่อ โปรแกรมรวบรวมเอกสารเว็บตรวจพบว่า รายการเว็บไซต์ที่ยังไม่ได้เข้าถึง ว่างเปล่าเมื่อต้องการ URL ที่จะทำการพิจารณา กระบวนการค้นหาจะหยุดชั่วคราว และทำการร้องขอบริการ Google SOAP Search API ด้วยคำสำคัญและกำหนด URL ตั้งต้นในฐานข้อมูลเพื่อให้โปรแกรมรวบรวมเอกสารเว็บสามารถเริ่มต้นการทำงานอีกครั้งได้

บางครั้งโปรแกรมรวบรวมเอกสารเว็บอาจเข้าสู่ ปัญหาสไปเดอร์แทรป (spider trap) ที่เกิดจากการเพิ่มจำนวนของ URL ที่อ้างอิงถึงเอกสารเว็บเดียวกัน หนทางหนึ่งที่จะแก้ไขปัญหาดังกล่าวคือ การจำกัดจำนวนของเอกสารเว็บที่โปรแกรมรวบรวมเอกสารเว็บจะทำการพิจารณาจากโดเมนหนึ่ง ซึ่งอาจมีผลกระทบทำให้วงของการค้นหาแคบลง

3.2.2 การร้องขอบริการจาก Google SOAP Search API

ในส่วนของ Google SOAP Search API จะเป็นการพัฒนาโปรแกรมเพื่อติดต่อกับ Google เพื่อค้นหาเว็บเพจจากคำสำคัญ โดยจะได้รายชื่อเว็บเพจที่ Google ทำการค้นหากลับมาแล้วนำรายชื่อเว็บไซต์มาเก็บลงในฐานข้อมูลเพื่อให้โปรแกรมรวบรวมเอกสารเว็บทำงานต่อไป โดยในการพัฒนาโปรแกรมในส่วนนี้จำเป็นต้องอาศัยมอดูล google ซึ่งการทำงานจะเหมือนกับการค้นหาในเว็บไซต์ของ Google การนำ Google SOAP Search API มาใช้งานต้องมีขั้นตอน ดังนี้

1. ดาวน์โหลด Google SOAP Search API ซึ่งจะประกอบไปด้วยเอกสารและตัวอย่างโปรแกรมที่มีทั้งโปรแกรมที่พัฒนาด้วยภาษาจาวา และ .NET และมีไฟล์ WDSL สำหรับรองรับการพัฒนาโปรแกรมบนแพลตฟอร์มต่างๆที่สนับสนุนการให้บริการเว็บ
2. สร้างบัญชีของ Google เพื่อที่เข้าใช้บริการในส่วนของ Google SOAP Search API ซึ่งเมื่อสร้างบัญชีแล้ว Google จะให้ license key เพื่อนำไปใช้พัฒนาโปรแกรมต่อไป โดยทั่วไป Google สามารถให้บริการการค้นหาเว็บเพจจากคำสำคัญต่อ 1 license key ได้เพียง 1000 คำต่อวัน

3. พัฒนาโปรแกรมโดยการใช้ license key ที่ได้รับ โดยในโปรแกรมจำเป็นต้องมีการระบุค่า license key ที่ได้รับมาในการค้นหาแต่ละครั้ง

เมื่อได้ license key มาจะนำมาพัฒนาโปรแกรมเพื่อติดต่อกับ google โดยโปรแกรมที่พัฒนาจะใช้คลาสและฟังก์ชันต่างๆของมอดูล google มาประกอบกัน
คลาสต่างๆของโมดูล Google มีดังนี้

1) SearchResult จะเก็บผลลัพธ์ต่างๆจากการค้นหา ซึ่งจะประกอบด้วยตัวแปรต่างๆ ดังนี้

- cachedSize : ขนาดแคชของผลลัพธ์ (KB)
- directoryCategory : รายละเอียดประเภทของไดเรกทอรีที่เปิดอยู่
- directoryTitle : titleของผลลัพธ์ของไดเรกทอรีที่เปิดอยู่
- hostName : ใช้เมื่อมีการกรองข้อมูล
- relateInformationPresent : คือคำที่เกี่ยวข้อง
- snippet : แสดงเนื้อหาของคำที่ค้นหา (HTML)
- summary : ประเด็นสำคัญสำหรับผลลัพธ์นั้น
- title : title (HTML)
- URL : URL

2) SearchResultsMetaData จะเป็นคลาสสำหรับ metadata เกี่ยวกับผลที่ได้จากคำที่ค้นหา ซึ่งประกอบด้วยตัวแปรต่างๆ ดังนี้

- directoryCategories : รายชื่อของประเภทสำหรับผลลัพธ์จากการค้นหา
- documentFiltering : เป็นตัวแปรที่บอกว่ามีการกรองหน้าซ้ำหรือไม่
- endIndex : อินเด็กซ์ของผลลัพธ์สุดท้ายที่ส่งค่ามา
- estimatedTotalResultsCount : จำนวนของเว็บเพจจากคำที่ค้นหาโดยประมาณ
- estimateIsExact : เป็นตัวแปรที่บอกว่าค่า estimatedTotalResultsCount เป็นค่าที่ถูกต้องหรือไม่
- searchComments : รายละเอียดที่คนทั่วไปสามารถเข้าใจได้ง่าย
- searchQuery : คำตั้งต้นของการค้นหา
- searchTime : เวลาทั้งหมดของการค้นหา (วินาที)
- searchTips : รายละเอียดเกี่ยวกับวิธีการใช้ Google
- startIndex : อินเด็กซ์ของผลลัพธ์เริ่มแรกที่ส่งค่ามา

3) **SearchReturnValue** ผลจากการค้นหาแบบสมบูรณ์ ซึ่งมีตัวแปรต่างๆ ดังนี้

- meta : SearchResultsMetaData instance สำหรับการค้นหา
- results : รายชื่อของ SearchResult object สำหรับการค้นหา

และฟังก์ชันที่สำคัญที่นำมาพัฒนาโปรแกรมมีดังนี้

1) **doGoogleSearch** เป็นฟังก์ชันที่ใช้ในการค้นหา Google โดยใช้ SOAP API และ Google จะทำการส่งค่ากลับมา โดยมีการใช้งานดังนี้

doGoogleSearch(q, start=0, maxResults=10, filter=1, restrict='', safeSearch=0, language='', inputencoding='', outputencoding='', license_key=None, http_proxy=None)

ฟังก์ชันนี้จำเป็นต้องใช้ license key และควรใส่ทุกครั้งที่มีการเรียกใช้ฟังก์ชันนี้ หรือตั้งค่าให้เป็น global ทุกอย่างที่สามารถทำได้ในเว็บไซต์ Google สามารถทำได้เช่นเดียวกันกับคำที่จะค้นหาในฟังก์ชันนี้ และสามารถใช้พารามิเตอร์ start และ maxResults เพื่อให้ได้ผลจำนวนของเว็บเพจหลายเว็บเพจ

หมายเหตุ maxResults สามารถกำหนดได้มีค่ามากสุดเท่ากับ 10

<u>พารามิเตอร์</u>

q – คำที่ใช้ในการค้นหา มีชนิดเป็นสตริง

start – อินเด็กซ์ของผลลัพธ์ตัวแรก มีค่าดีฟอลต์เป็น 0 (ไม่จำเป็นต้องใส่) มีชนิดเป็นinteger

maxResults – ค่าผลลัพธ์สูงสุดที่ Google ส่งค่ากลับมา (ไม่จำเป็นต้องใส่) มีชนิดเป็น integer

filter – ค่าแฟลกที่ตั้งไว้เพื่อกลั่นกรองเว็บเพจที่ซ้ำกัน (ไม่จำเป็นต้องใส่) มีชนิดเป็น integer

restrict – ข้อจำกัดในการค้นโดยใช้ประเทศหรือหัวข้อเป็นตัวจำกัด (ไม่จำเป็นต้องใส่) มีชนิดเป็นสตริง เช่น U.S. Gorvernment (unclesam), Linux (linux), Macintosh (mac) และ FreeBSD (bsd)

safeSearch – ค่าแฟลกที่ตั้งไว้เพื่อกลั่นกรองเนื้อหา (ไม่จำเป็นต้องใส่) มีชนิดเป็น integer

language – เป็นข้อจำกัดทางภาษา มีชนิดเป็นสตริง

inputencoding – การเข้ารหัสของอินพุท (ไม่จำเป็นต้องใส่) มีชนิดเป็นสตริง เช่น UTF-8

outputencoding – การเข้ารหัสของเอาท์พุท (ไม่จำเป็นต้องใส่) มีชนิดเป็นสตริง

license_key – license key ของ Google API ที่ใช้ (ไม่จำเป็นต้องใส่) มีชนิดเป็นสตริง

http_proxy – HTTP proxy ที่ใช้ในการติดต่อกับ Google (ไม่จำเป็นต้องใส่) มีชนิดเป็นสตริง

ค่าที่ Google ส่งกลับมาเป็นผลลัพธ์จากการค้นหาที่ถูกเก็บเป็นออบเจ็กต์ มีชนิดเป็น SearchReturnValue

ตัวอย่างการใช้งานเพื่อให้แสดงลิงค์ของเว็บเพจจากการค้นหา 'project'

```
google.LICENSE_KEY = 'IQiYLq9QFHKl/m1m1MgRSL07xdOj1UL0'
data = google.doGoogleSearch('project',0)
for result in data.results:
        print result.URL
```

ในตอนแรกจะต้องทำการกำหนดค่า license key เพื่อนำไปใช้ในการค้นหาต่อไป จากนั้นทำการติดต่อกับ Google เพื่อให้ Google ค้นหาคำว่า 'project' ให้โดยใช้ฟังก์ชัน doGoogleSearch จากโปรแกรมตัวอย่างด้านบนจะให้โปรแกรมแสดงลิงค์ของเว็บเพจที่ได้จากการค้นหา 10 เว็บเพจ แต่ถ้าต้องการให้โปรแกรมแสดงลิงค์เว็บเพจทั้งหมดที่ได้จากการค้นหา จำเป็นทำเป็นลูปและกำหนดอินเด็กซ์ของผลลัพธ์ตัวแรกให้มีค่าเท่ากับอินเด็กซ์ตัวสุดท้ายของการค้นหาในรอบที่แล้ว

```
data = google.doGoogleSearch(('project',data.meta.endIndex)
```

3.2.3 การร้องขอเอกสารเว็บจากเครื่องที่ให้บริการ

ในการดึงข้อมูลจากหน้าเว็บเพจซึ่งเครื่องที่ขอใช้บริการ (HTTP client) จะทำการส่งคำร้องขอ (HTTP request) สำหรับเว็บเพจและอ่านค่าจากคำตอบรับ (HTTP response) โดยเครื่องที่ขอใช้บริการต้องมีค่าเวลาสิ้นสุด (timeout) ซึ่งเป็นการกำหนดช่วงเวลา เพื่อให้แน่ใจว่าไม่ใช้เวลาโดยไม่จำเป็นบนเครื่องที่ให้บริการที่ล่าช้าหรือในการอ่านหน้าเว็บเพจที่มีขนาดใหญ่

กระบวนการตรวจสอบความผิดพลาดมีความสำคัญมาก ในขณะที่จะต้องทำการดึงเว็บเพจจากเครื่องที่ให้บริการจำนวนมากด้วยโปรแกรมแบบเดียวกัน

โครงงานนี้ได้ใช้ภาษาไพธอน (Python) ซึ่งเป็นภาษาระดับสูง สามารถรองรับการทำงานเพื่อเข้าสู่เครื่องที่ให้บริการและร้องขอ เอกสารเว็บจาก URL มาเพื่อประมวลผลได้ง่ายและมีประสิทธิภาพสูงและเลือกใช้ไลเบอรี่มาตรฐาน urllib2 และ httplib เพื่อใช้ในการร้องขอเอกสารเว็บและเชื่อมต่อกับเครื่องที่ให้บริการตามลำดับ

```
>>> import urllib2
>>> import httplib
>>> myURL = "http://www.ce.kmitl.ac.th"
>>> tempFile = urllib2.urlopen(urllib2.Request(myURL))
>>> text = tempFile.read()
>>> print text

<HTML><HEAD><TITLE>ภาควิชาวิศวกรรมคอมพิวเตอร์</TITLE>

<meta "robots" content="index,nofolow">

<meta "robots" content="noindex,nofolow">
```

รูปที่ 3.3 ตัวอย่างการเรียกใช้ Library มาตรฐาน urllib2 และ httplib ของภาษา Python

3.2.4 การค้นหาลิงค์ของเว็บเพจและภาพภายในเอกสารเว็บ

เมื่อเว็บเพจถูกดึงจะต้องทำการวิเคราะห์เนื้อหาในเว็บเพจ เพื่อดึงข้อมูลที่ต้องการ ซึ่งกระบวนการวิเคราะห์เนื้อหาในเว็บเพจจะมีกระบวนการบนเอกสารเว็บ (HTML content)

กระบวนการวิเคราะห์เนื้อหาในเอกสารเว็บมีอิสระกว้างขวางขึ้นกับภาษาคอมพิวเตอร์ที่ใช้ ซึ่งมีการจัดให้ฟังก์ชันที่ง่ายในการระบุถึง HTML tag และทำการแยกส่วน hyperlink URL ออกมาจากเว็บเพจ ซึ่งเราสามารถใช้ตัววิเคราะห์เนื้อหาในการค้นหา anchor tag ซึ่งเป็น tag ที่เก็บ hyperlink และเก็บค่าที่อยู่ใน attribute href ได้ อย่างไรก็ตามเราจำเป็นต้องทำการแปลง relative URL ไปเป็น URL ที่สมบูรณ์ โดยใช้ URL พื้นฐานของเว็บเพจนั้นจากที่ได้รับมา

URL ที่แตกต่างกันแต่อ้างถึงเว็บเพจเดียวกัน สามารถแปลงเป็น URL เดียวกันได้ โดยมีขั้นตอนบางส่วนดังนี้

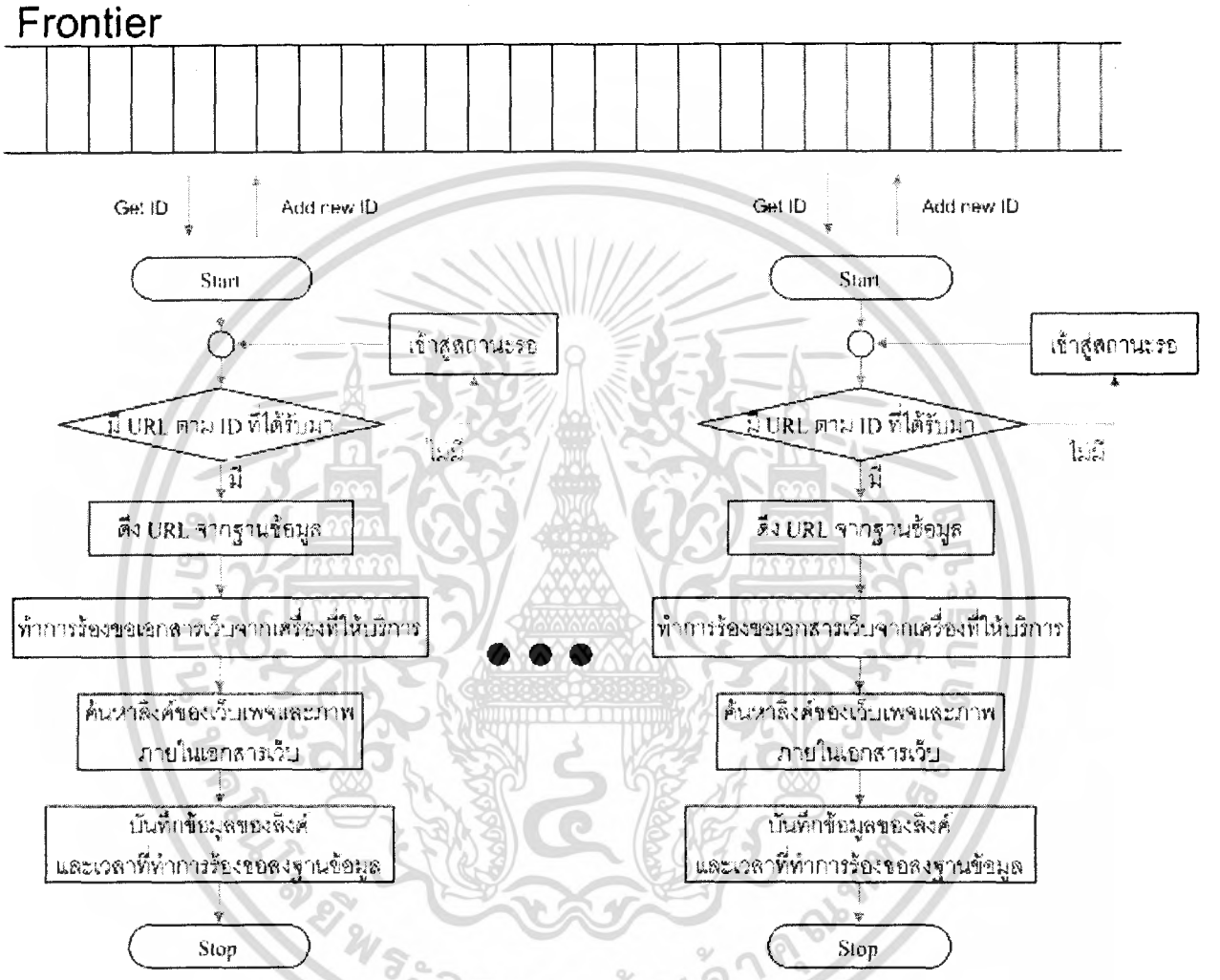
- ทำการกลับอักษรที่แสดง protocol และ hostname เป็นอักษรตัวเล็กเช่น HTTP://WWW.ce.KMITL.ac.th กลับอักษรเป็น http://www.ce.kmitl.ac.th
- ทำการลบส่วน 'anchor' หรือ 'reference' ออกจาก URL สำหรับลิงค์ที่แสดงถึงส่วนของเว็บเพจเดียวกัน (Infrapage link)เช่น http://myspiders.biz.uiowa.edu/faq.html#what ลดลงเป็น http://myspiders.biz.uiowa.edu/faq.html
- ต้องทำการเพิ่มชื่อโดเมนเนม (domain name) ให้กับลิงค์ที่แสดงเว็บเพจที่อยู่ในโดเมนเนมเดียวกัน (Intrasystem link)เช่น .... เมื่อทำการดึงลิงค์ออกมาจาก tag จะได้ file.html ต้องทำการปรับปรุงเป็น http://www.kmitl.ac.th/file.html
- สำหรับบาง URL ต้องทำการเพิ่ม '/' ตอนท้ายเช่น http://www.sanook.com และ http://www.sanook.com/ จะต้องทำการแปลงไปเป็น URL เดียวกันโดยต้องตัดสินใจว่าจะทำการเพิ่ม '/' ตอนท้ายหรือไม่

โดยกระบวนการดังกล่าวอาจจะต้องใช้ร่วมกันในบางกรณี โดยจะต้องทำการพิจารณาเพื่อลดปัญหาต่างๆ เช่น สไปเดอร์แทรป โดยปกติในกรณีที่เกิดสไปเดอร์แทรปขนาดของ URL จะเพิ่มขึ้นเรื่อยๆ ดังนั้นสามารถจำกัดขนาดของ URL โดยออกแบบให้เท่ากับ 256 ตัวอักษรเพื่อลดปัญหาที่เกิดขึ้น

การเขียนโปรแกรมที่ทำการดึงส่วนของลิงค์ออกมาจากเอกสารเว็บนั้นภาษาไพธอน (Python) สามารถทำได้โดยการเรียกใช้ไลเบอรี่มาตรฐาน htmllib ที่สามารถดึงลิงค์ที่อยู่ใน tag <a> และสามารถทำการเขียนทับการทำงาน (Override function) ให้ทำการพิจารณา tag <img> ซึ่งเก็บข้อมูลรูปภาพและทำการดึง URL ของรูปภาพจาก attribute src ได้เช่นเดียวกัน

3.2.5 โปรแกรมรวบรวมเอกสารเว็บแบบมัลติเทรด (Multi-threaded Crawler)

ลักษณะของรอบการค้นหาในลักษณะลำดับ (sequential) จำเป็นต้องใช้เวลามากจึงเปลี่ยนการทำงานด้วยกระบวนการมัลติเทรดดิ้ง (Multi-threading) ซึ่งแต่ละเทรด (thread) ตามแต่ละรอบของการค้นหาสามารถเพิ่มความเร็ว และประสิทธิภาพของการใช้ช่องเครือข่าย (bandwidth)



รูปที่ 3.4 การทำงานของ Multi-threaded Crawler

รูปที่ 3.4 แสดงการทำงานของโปรแกรมรวบรวมเอกสารเว็บแบบมัลติเทรด (Multi-threaded Crawler) ที่ปรับปรุงจากรูปที่ 3.2 โดยแต่ละเทรดเริ่มการทำงาน รับค่าระบุถึง URL ในฐานข้อมูล (ID) เพื่อทำการดึง URL ไปสู่การค้นหา หลังจากนั้นจะทำการทำงานแบบเดียวกับโปรแกรมรวบรวมเอกสารเว็บแบบเป็นลำดับปกติ เมื่อเสร็จสิ้นเทรดจะทำการเพิ่ม ID ใหม่เข้าสู่รายการเว็บไซต์ที่ยังไม่ได้เข้าถึง (frontier) และสิ้นสุดการทำงาน หลังจากนั้นโปรแกรมจะทำการสร้างเทรดใหม่ขึ้นมาทำงานทดแทนตลอดเวลาโดยมีจำนวนของเทรดตั้งต้นออกแบบให้เท่ากับ 128 เทรด

ความแตกต่างระหว่างการค้นหาแบบมัลติเทรดและการค้นหาแบบเป็นลำดับปกติ คือ ในขณะที่เทรดพบว่า frontier ว่างเปล่าเทรดจะไม่ร้องขอบริการของ Google SOAP Search API แบบอัตโนมัติทันที เพราะยังมีโอกาสที่เทรดอื่นอาจจะทำการดึงเว็บเพจหรือทำการเพิ่ม ID ของ URL ใหม่ในอนาคตอันใกล้ โดยจะให้เทรดนั้นอยู่ในสถานะรอ (waiting) จนกว่าทุกๆ เทรดจะเข้าสู่สถานะรอ ซึ่งหมายความว่าไม่มีเทรดใดทำการเพิ่ม ID ของ URL ใหม่ในอนาคต โปรแกรมรวบรวมเอกสารเว็บจะทำการร้องขอบริการของ Google SOAP Search API แบบอัตโนมัติทันที เพื่อให้เทรดที่อยู่ในสถานะรอ ทำงานต่อไปได้ตามปรกติ

3.3 โปรแกรมตรวจสอบภาพอนาจาร (Pornographic detection)

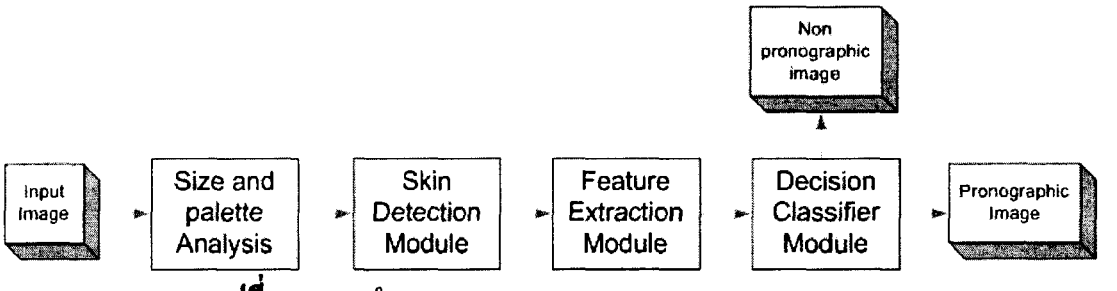
การตรวจหาภาพอนาจารภายในรูปภาพจะทำการระบุตำแหน่งของพิกเซลของผิวหนังและรวมกลุ่มไว้ด้วยกันเป็นกลุ่มของผิวหนัง โดยใช้รายละเอียดของสี, พื้นผิว และขอบ คุณลักษณะต่างๆถูกดึงมาจากกลุ่มของผิวหนังและใช้เป็นอินพุทเพื่อใส่เข้าไปในการจัดกลุ่ม

ความสัมพันธ์ระหว่างเปอร์เซ็นต์ของสีผิวกับรูปอนาจารมีค่อนข้างมาก ในขั้นตอนแรกต้องมีการแบ่งกลุ่มของสีผิว แต่การจำแนกสีของผิวหนังอาจมีความผิดพลาดได้ เนื่องจาก

1) ภาพมีคุณภาพต่ำ เช่น ภาพที่มี contrast ต่ำ
2) ภาพที่มีวัตถุที่มีสีใกล้เคียงกับสีผิวมนุษย์
3) ภาพผิวหนังที่มีสีผิวจางเนื่องจากการส่องสว่างและการสะท้อนของแสง

การตรวจสอบภาพอนาจารมี 4 ขั้นตอนหลักๆ ดังรูปที่ 3.5 คือ

1. Size and palette Analysis เป็นการกรองภาพแบบพื้นฐานที่สุด เช่น ภาพมีขนาดเล็กกว่าที่เราตั้งเอาไว้ และภาพที่มีจำนวนสีไม่มาก (น้อยกว่า 50 พิกเซล) เนื่องจากคุณสมบัติไม่เหมือนกับภาพอนาจาร
2. Skin Detection ในขั้นตอนนี้เราจะพิจารณาจากสี, พื้นผิวและขอบ พิจารณาพื้นผิวที่เชื่อมต่อกัน
3. Feature Extraction การดึงคุณสมบัติ 6 อย่างจากข้อมูลภาพ
4. Decision Classifier เป็นการจัดประเภทของข้อมูลภาพ



รูปที่ 3.5 การทำงานของการตรวจสอบภาพอนาจาร

3.3.1 การกำจัดภาพขนาดเล็กกว่ากำหนด (Size and palette Analysis)

จากสมมติฐานภาพอนาจารควรมีขนาดใหญ่และใช้พื้นที่ส่วนใหญ่ของหน้าเอกสารเว็บ ดังนั้นเพื่อจำกัดการตรวจสอบและเพิ่มประสิทธิภาพในการตรวจสอบในขั้นต้น ทำการตัดภาพที่มีขนาดเล็กว่ากำหนด ออกแบบให้เท่ากับ 50x50 พิกเซล ภาพที่มีขนาดเล็กกว่าที่กำหนดจะไม่ถูกนำมาพิจารณา

3.3.2 การตรวจหาสีผิวมนุษย์ (skin detection module)

ในการตรวจหาสีผิวมนุษย์ที่ออกแบบมี 4 ขั้นตอนย่อยดังนี้

1. การตรวจหาพิกเซลในช่วงของสีผิว (Skin Tone Color Detection) จะพิจารณาจากค่า RGB ที่ถูกแปลงเป็นค่า YCbCr ว่าพิกเซลนั้นมีค่าอยู่ในช่วงสีผิวหรือไม่
2. การตรวจหาขอบเขตพื้นที่ของพิกเซลที่มีค่าอยู่ในช่วงสีผิว (Skin Region Expansion) จากการพิจารณาข้อที่ 1 อาจมีการรวมถึงพิกเซลข้างเคียงที่มีสีอยู่ในช่วงของสีผิว
3. การแบ่งพื้นที่ของสีผิวด้วยเส้นขอบ (Skin Region Segmentation) การตรวจหาขอบเขตจากข้อที่ 2 จะมีการพิจารณาร่วมกับกระบวนการตรวจหาขอบของโซเบล (Sobel) พิกเซลที่อยู่บนขอบจะไม่ถูกตรวจจับว่าเป็นสีผิว การทำเช่นนี้เพื่อให้แน่ใจว่าเป็นพื้นที่ของผิวจริง
4. การกำจัดพื้นที่ของสีผิวที่มีขนาดเล็กกว่าที่กำหนดไว้ (Skin Blob Detection) กำหนดไว้ที่ 50x50 พิกเซล

3.3.3 การตรวจสอบคุณสมบัติของภาพ (Feature Extraction Module)

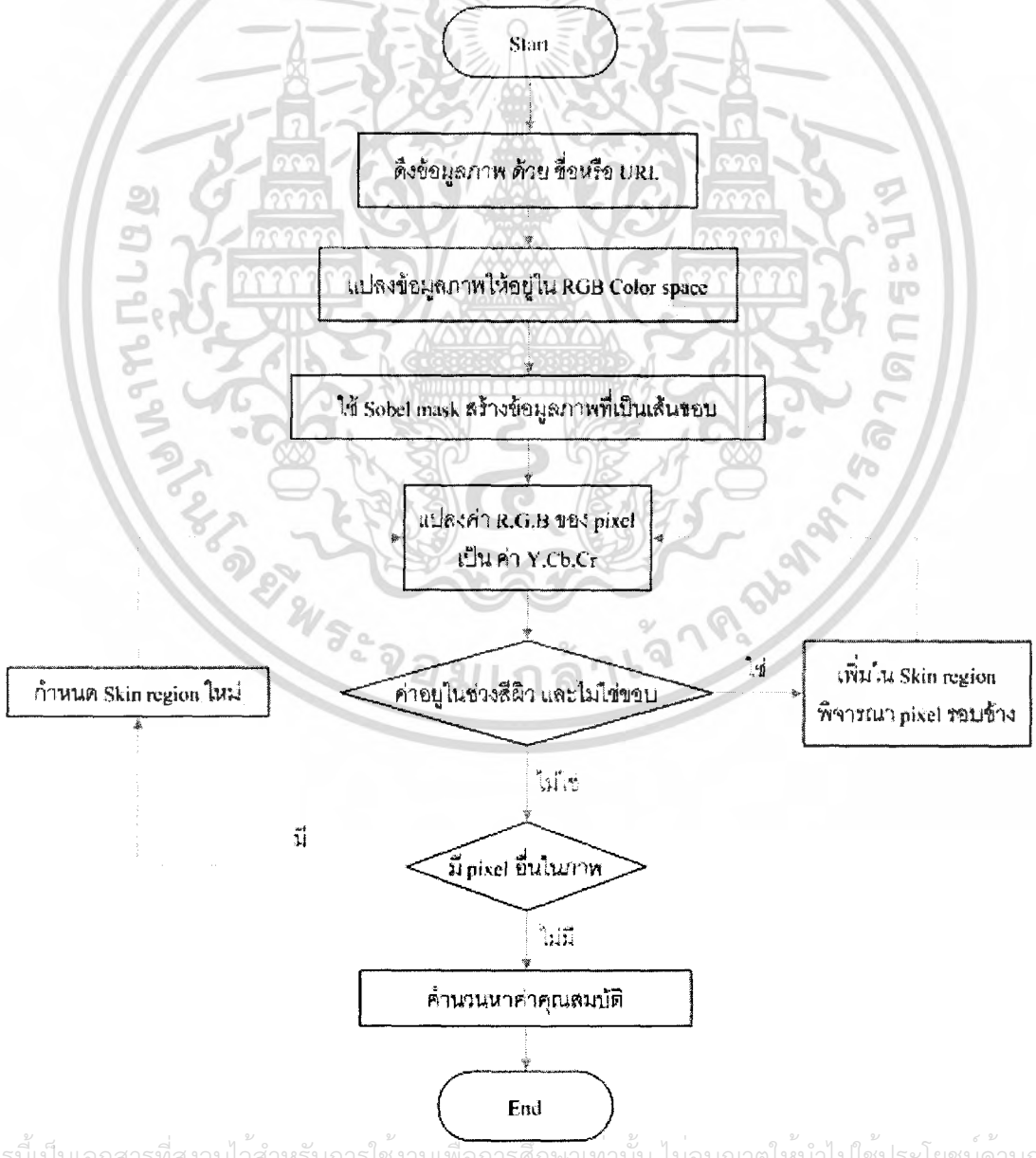
ภาพที่ถูกตรวจสอบว่าเป็นภาพอนาจาร จะถูกพิจารณาคุณสมบัติ 6 ข้อ ดังตารางที่ 3.1

ลำดับ	คุณสมบัติ	รายละเอียด
1.	เปอร์เซ็นต์ของสีผิวในภาพ	อัตราของพิกเซลที่มีค่าอยู่ในช่วงของสีผิวทั้งหมดเทียบขนาดของภาพ
2.	เปอร์เซ็นต์ของพื้นที่สีผิวที่มีขนาดใหญ่ที่สุด	อัตราของพิกเซลในพื้นที่ที่มีสีผิวมากที่สุดเทียบกับขนาดของภาพ
3.	จำนวนชิ้นส่วนของพื้นที่ผิว	จำนวนของพื้นที่ที่มีสีผิวในภาพ
4.	อัตราความสูงของพื้นที่ผิว	ขนาดความสูงของพื้นที่ที่มีสีผิวขนาดใหญ่ที่สุดต่อขนาดความสูงของภาพ
5.	อัตราความกว้างของพื้นที่ผิว	ขนาดความกว้างของพื้นที่ที่มีสีผิวขนาดใหญ่ที่สุดต่อขนาดความกว้างของภาพ
6.	อัตราส่วนของพื้นที่ผิวขนาดใหญ่เทียบกับจำนวนพื้นที่ผิวทั้งหมด	อัตราส่วนของพื้นที่ที่มีสีผิวขนาดใหญ่ที่สุดในภาพเทียบกับขนาดพื้นที่ที่มีสีผิวทั้งหมดในภาพ

ตารางที่ 3.1 คุณสมบัติ 6 ประการที่ได้จากภาพ

คุณสมบัติทั้ง 6 ข้อสามารถใช้จำแนกภาพอนาจารได้ โดยภาพที่ไม่มีกลุ่มของสีผิวจะถูกมองว่าไม่เป็นภาพอนาจาร ซึ่งไม่มีคุณสมบัติข้อที่ 1 จากสมมติฐานภาพอนาจารจะต้องประกอบด้วยกลุ่มของสีผิวขนาดใหญ่ต่อเนื่องกัน ซึ่งสามารถพิจารณาได้จากคุณสมบัติข้อที่ 2 และ 3 ที่แสดงถึงอัตราส่วนของพื้นที่ผิวที่มีขนาดใหญ่ที่สุดและจำนวนชิ้นส่วนของพื้นที่ผิว ภาพที่มีจำนวนของพื้นที่ผิวมากอาจจะไม่ใช่ภาพอนาจาร ความสัมพันธ์ระหว่างขนาดและรูปร่างของกลุ่มสีผิว ถ้าเป็นภาพอนาจารจะมีกลุ่มสีผิวขนาดใหญ่ คุณสมบัติข้อ 5 และ 6 เป็นความสัมพันธ์ของอัตราส่วนของความกว้างและความสูงของกลุ่มสีผิวต่อความกว้างและความสูงของภาพ มีแนวโน้มว่ากลุ่มสีผิวจะครอบคลุมพื้นที่ส่วนใหญ่ของภาพ

โครงงานนี้ได้จัดทำโปรแกรมตรวจสอบภาพอนาจารที่สามารถวิเคราะห์คุณสมบัติของภาพได้ โดยมีขั้นตอนการทำงานดังรูปที่ 3.6

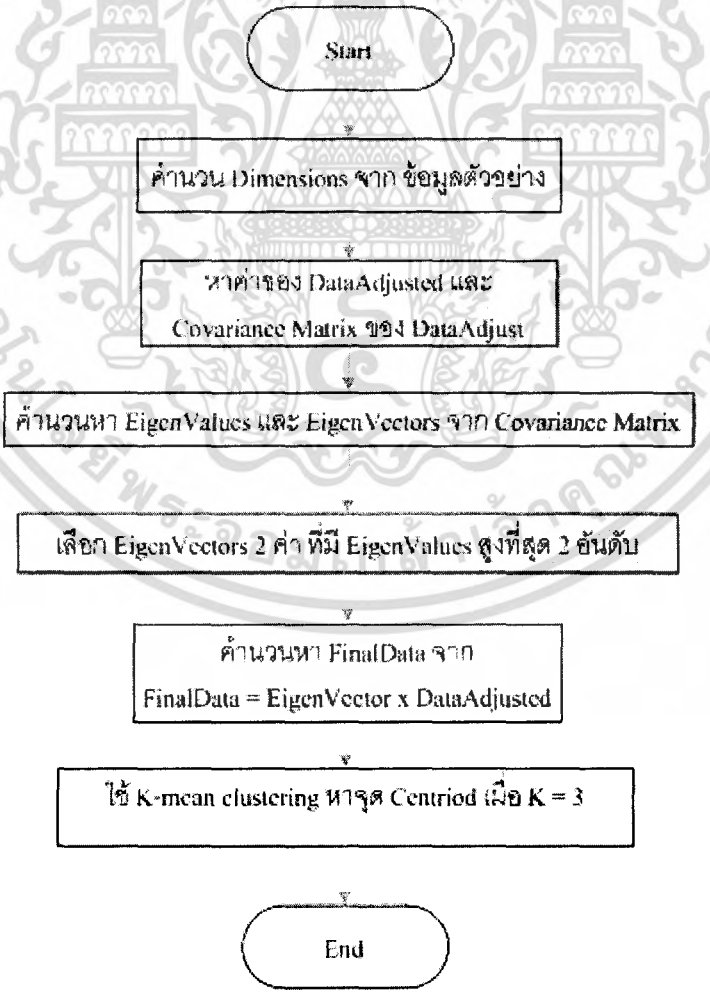


รูปที่ 3.6 การทำงานของโปรแกรมการตรวจสอบภาพอนาจาร

3.3.4 การจัดประเภทของภาพด้วยคุณสมบัติ (Decision Classifier)

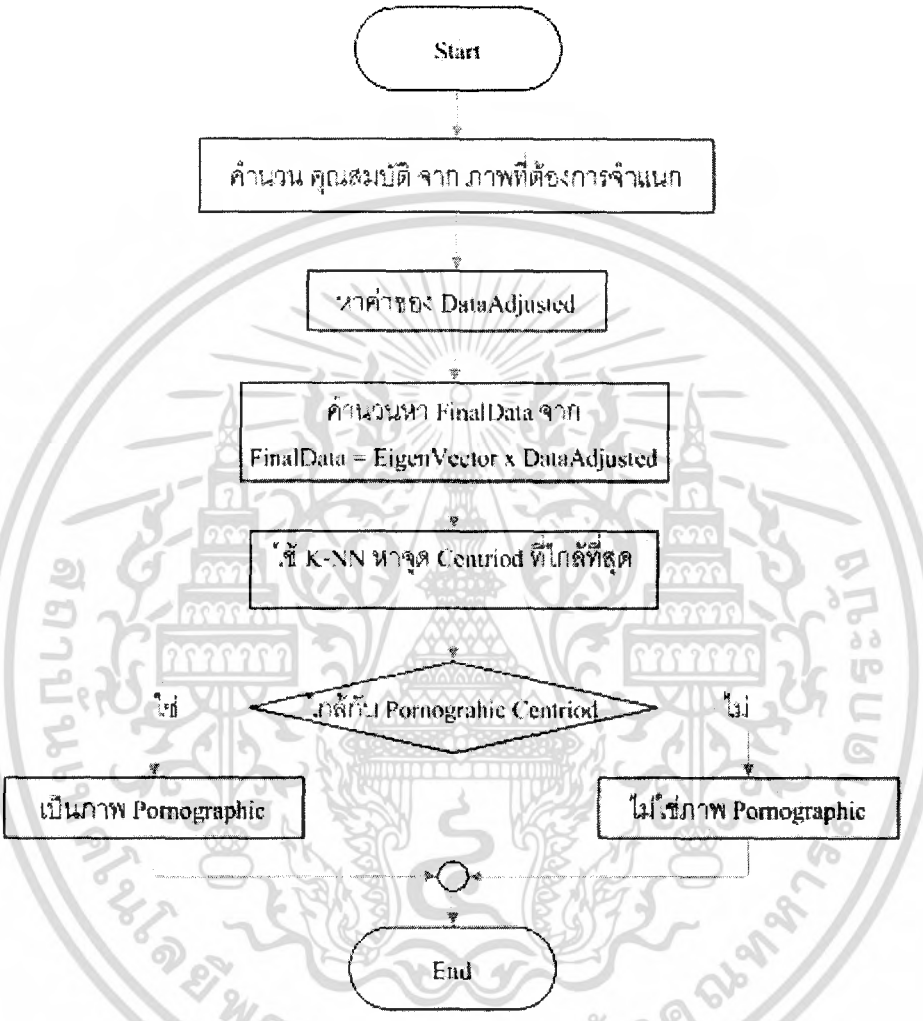
เมื่อพิจารณาคุณสมบัติของภาพทั้ง 6 ข้อ เราสามารถจำแนกประเภทของภาพได้ 3 กลุ่ม ได้แก่ ภาพอนาจาร , ภาพที่ไม่ใช่ภาพอนาจารแต่มีส่วนประกอบของสีผิว และภาพที่ไม่มีส่วนประกอบของสีผิว เพื่อให้สามารถพิจารณาคุณสมบัติที่มีความสำคัญมากที่สุดจำเป็นต้องทำการรวมคุณสมบัติและลดคุณสมบัติบางประการโดยไม่ทำให้ข้อมูลสูญหายมากเกินไปด้วยหลักการการวิเคราะห์ส่วนประกอบที่ต้องทำการคำนวณหาไอแกนเวกเตอร์และไอแกนแวลูส์ของชุดข้อมูลตัวอย่าง แล้วทำการหาจุดเซ็นทรอยด์ 3 จุดที่แสดงถึงข้อมูลภาพ 3 ประเภทด้วยหลักการของ K-mean Clustering และเมื่อต้องการจำแนกภาพสามารถทำได้โดยการคำนวณหาระยะทางที่ใกล้ที่สุดเทียบกับจุดเซ็นทรอยด์ด้วยหลักการของ K-Nearest Neighbor (KNN)

โครงงานนี้ได้จัดทำโปรแกรมจัดประเภทของภาพด้วยคุณสมบัติ ที่สามารถจำแนกประเภทของภาพด้วยคุณสมบัติของภาพได้ โดยมีขั้นตอนการคำนวณหาไอแกนเวกเตอร์และไอแกนแวลูส์ของชุดข้อมูลตัวอย่าง แล้วทำการหาจุดเซ็นทรอยด์ 3 จุดที่แสดงถึงข้อมูลภาพ 3 ประเภทด้วยหลักการของ K-mean Clustering ดังรูปที่ 3.7



รูปที่ 3.7 การทำงานของการคำนวณหาไอแกนเวกเตอร์และไอแกนแวลูส์ของชุดข้อมูลตัวอย่างและ การหาจุดเซ็นทรอยด์ 3 จุด

เมื่อคำนวณหาไอแกนเวกเตอร์และไอแกนแวลูส์ของชุดข้อมูลตัวอย่าง แล้วทำการหาจุดเซ็นทรอยด์ 3 จุด ได้จะกระบวนการข้างต้นแล้ว เมื่อต้องการจำแนกภาพสามารถทำได้โดยการคำนวณหาระยะทางที่ใกล้ที่สุดเทียบกับจุดเซ็นทรอยด์ด้วยหลักการของ K-Nearest Neighbor (KNN) มีขั้นตอนการทำงานดังรูปที่ 3.8



รูปที่ 3.8 การจำแนกภาพ

3.4 โปรแกรมการวิเคราะห์เนื้อหาในเว็บไซต์ (Text Detection)

การวิเคราะห์เนื้อหาในเว็บไซต์นั้นจะใช้กระบวนการดึงส่วนที่เป็นเอกสาร HTML ของเว็บเพจมาทำการตรวจสอบเพื่อพิจารณาโครงสร้างและคุณสมบัติต่างๆภายในหน้าเว็บเพจนั้นๆ ซึ่งคุณสมบัติที่นำมาพิจารณามีทั้งสิ้น 10 คุณสมบัติ ดังนี้

1) จำนวน IMG Tag
2) จำนวน META Tag
3) จำนวนคำที่ไม่เหมาะสมใน TITLE Tag และ META Tag

4) จำนวน SCRIPT Tag

5) จำนวนลิงค์ทั้งหมด

6) จำนวนลิงค์ที่เป็นรูปภาพ

7) จำนวนลิงค์ที่เป็นข้อความ

8) จำนวนลิงค์ที่ไม่ใช่ลิงค์ภายใน

9) จำนวนลิงค์ที่เป็นลิงค์ภายใน

10) จำนวน PARAM Tag

เว็บไซต์อนาจารและเว็บไซต์ไม่อนาจารจะมีคุณสมบัติทั้ง 10 แตกต่างกันทำให้สามารถแยกแยะเว็บไซต์อนาจารจากเว็บไซต์ไม่อนาจารได้ โดยการตรวจสอบคุณสมบัติทั้ง 10 นั้นจำเป็นต้องใช้ Regular Expression เข้ามาช่วยเพื่อให้ค้นหามีประสิทธิภาพมากขึ้น

Regular Expression ในภาษาไพธอนใช้มอดูล re ซึ่งในการพัฒนาโปรแกรมในส่วนนี้จะมีการใช้ฟังก์ชันและแฟล็กต่างๆ ดังนี้

- re.search เป็นฟังก์ชันที่ทำการค้นหาโดยจะส่งค่ากลับเป็น none ถ้าไม่พบ
- re.findall เป็นฟังก์ชันที่ทำการค้นหาแต่จะทำการค้นหาทั้งหน้าเอกสารและจะส่งค่ากลับมาเป็นอาเรย์ของทูเพิล ซึ่งภายในทูเพิลจะเป็นสตริงที่มีคำที่ค้นหาประกอบอยู่
- re.I เป็นค่าแฟล็กภายในมอดูล re ซึ่งหมายถึงการพิจารณาอักษรตัวใหญ่และตัวเล็กเป็นตัวเดียวกัน

ตัวอย่างการค้นหาคำโดยใช้ Regular Expression ในโปรแกรม

```
img = "<((\s)?)+img"
find = re.findall(img,htmlSource,re.I)
if find:
   numimg = len(find)
else:
   numimg = 0
```

ตัวอย่างด้านบนเป็นการค้นหา IMG Tag ในเอกสารเว็บ(HTML Source) โดยคำที่ต้องการค้นหา(IMG Tag) จะอยู่ในรูปแบบของ Regular Expression เมื่อโปรแกรมพบ IMG Tag จะเก็บจำนวนที่พบ (len(find)) ในตัวแปรnumimg ถ้าไม่พบค่าตัวแปร numimg จะมีค่าเท่ากับ 0

ตัวอย่างคำที่ใช้ค้นหา ซึ่งอยู่ในรูปแบบของ Regular Expression

```
meta = "<((\s)?)+meta"
script = "<((\s)?)+script"
link = "<((\s)?)+a((\s)?)(.*)href"
```

```
linkimg = "<((\s)?)+a((\s)?)(.*)href(.*)>((\s)?)+<((\s)?)+img(.*)>
          ((\s)?)+<((\s)?)+/((\s)?)+a((\s)?)+>"
outlink = '(<a[\s]+)([^>]*)([\s]*href[\s]*=[\s]*[\"]?)([\s]*(http:\/\/))'
outlink1 = '(<a[\s]+)([^>]*)([\s]*href[\s]*=[\s]*[\"]?)([\s]*(https:\/\/))'
param = "<((\s)?)+param"
```

ตัวอย่างโปรแกรมในส่วนที่ค้นหาคำไม่เหมาะสมใน META Tag

```
findmeta = re.findall("<((\s)?)+meta",htmlSource,re.I)
for i in findmeta:
    if index<len(findmeta):
        for j in findmeta[index]:
            findd = re.findall("([^a-z0-9]*)([a-z0-9]+)([^a-z0-9]*)",j,re.I)
            for k in findd:
                pool = "^18$|^adulter$|(.*)amateurcouple(.*)|^anal$|..."
                a = pool.split("|")
                for m in range(len(a)):
                    searchpool = re.findall(a[m],k[1],re.I)
                    if searchpool:
                        poolword = poolword+len(searchpool)
```

จากโปรแกรมด้านบนจะทำการค้นหา META Tag ในเอกสารเว็บก่อน ผลที่ได้จากการค้นหาจะเป็นอาเรย์ของทูเพิล ภายในทูเพิลจะเป็นสตริงภายใน META Tag จึงต้องนำมาตัดเป็นคำเพื่อนำมาตรวจสอบต่อไป โดยการตัดคำนั้นจะพิจารณาจากอักขระพิเศษ เช่น ช่องว่าง เมื่อได้คำภายในแท็กจะนำมาตรวจสอบคำที่ไม่เหมาะสมภายในคำที่ตัดมา สุดท้ายนำผลลัพธ์ที่เป็นจำนวนของคำที่ไม่เหมาะสมมาเก็บไว้ในตัวแปร ซึ่งคำที่ไม่เหมาะสมที่นำมาทำการค้นหามีดังนี้

^adulter$|(.*)amateurcouple(.*)|^anal$|^anilingus(.*)|^anus$|^ass$|(.*)bdsm(.*)|(.*)blowjob(.*)|^bondage$|^boob(s)?$|^boobie(s)?$|(.*)borrachas(.*)|(.*)bulldoglist(.*)|^busty$|^cam(s)?$|(.*)cfnm(.*)|^cock(s)?$|(.*)creampie(.*)|^cum$|^cumshot(s)?$|^cunt$|(.*)cybersex(.*)|(.*)cybersexual(.*)|^dildo(s)?$|(.*)femdom(.*)|(.*)fuck(.*)|(.*)handjob(.*)|(.*)hentai(.*)|(.*)interracial(.*)|(.*)maledom(.*)|(.*)malestripper(.*)|(.*)masturbat(.*)|^mature(.*)|(.*)milf(.*)|^nude(.*)|^nudism(.*)|(.*)nudistas(.*)|^nudity(.*)|^oral(.*)|(.*)orgasm(.*)|^pantie(.*)|(.*)penis(.*)|^piss(.*)|(.*)porn(.*)|(.*)pued(.*)|(.*)purecfnm(.*)|^puss(.*)|^putas(.*)|^rimjob(.*)|^rimming(.*)|(.*)rubias(.*)|(.*)shemale(.*)|(.*)tetazas(.*)|(.*)tgp(.*)|^tit(s)?$|(.*)titworld(.*)|(.*)upskirt(.*)|(.*)xnxx

(.*)|(.*)xxx(.*)|(.*)yobt(.*)|(.*)ztod(.*)|(.*)gangbang(.*)|(.*)hardcore(.*)|^babe(s)?$|^amateur(s)?$|^DVDA$|^pussy$

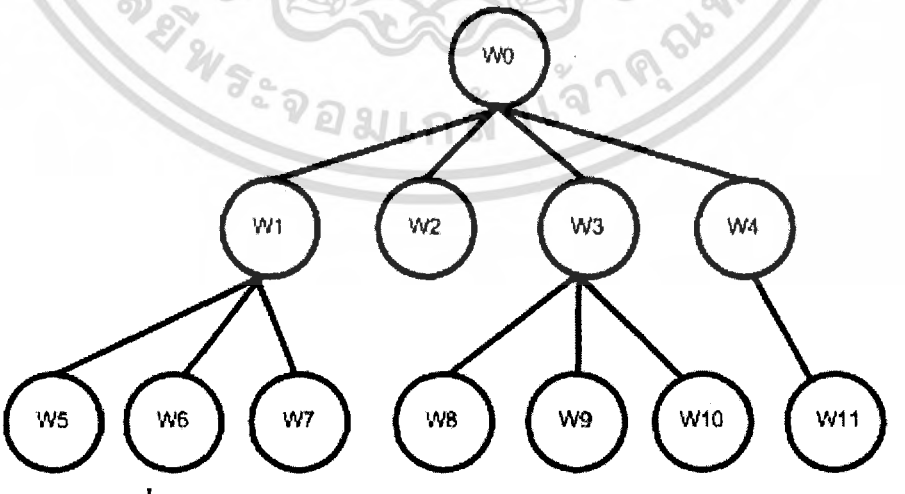
3.4.1 การจัดประเภทของเนื้อหาเว็บไซต์

เมื่อพิจารณาคุณสมบัติทั้ง 10 ประการแล้ว จะทำการจำแนกประเภทของเนื้อหาเว็บไซต์เป็น 2 ประเภท คือ เว็บไซต์อนาจารและเว็บไซต์ไม่อนาจาร โดยการผ่านหลักการการวิเคราะห์ส่วนประกอบ จะทำการรวมและลดคุณสมบัติให้เหลือเพียง 2 ประการซึ่งต้องทำการคำนวณหาไอแกนเวกเตอร์และไอแกนแวลูส์ของชุดข้อมูลตัวอย่าง จากนั้นเมื่อพล็อตจุดของข้อมูลตัวอย่างที่ผ่านกระบวนการวิเคราะห์ส่วนประกอบแล้วต้องทำการหาสมการที่สามารถแยกแยะข้อมูลเว็บไซต์อนาจารและเว็บไซต์ไม่อนาจารจากกันได้

ข้อมูลใหม่ที่รับมาจะนำมาผ่านกระบวนการการวิเคราะห์ส่วนประกอบ เมื่อนำข้อมูลนั้นมาพล็อตกราฟ ถ้าจุดข้อมูลอยู่เหนือสมการที่หามาได้จากกลุ่มตัวอย่างแสดงว่าเว็บไซต์นั้นไม่ใช่เว็บไซต์อนาจารแต่ถ้าจุดข้อมูลนั้นอยู่ใต้สมการแสดงว่าเว็บไซต์นั้นเป็นเว็บไซต์อนาจาร

3.5 โปรแกรมการคำนวณค่าน้ำหนักความไม่เหมาะสม

เราสามารถมองลักษณะของเว็บไซต์ต่างๆ ในรูปของทรี(tree) ได้โดยประกอบด้วย URL ของเว็บไซต์เป็นรูทโหนด (root node) และมีลิงค์ของเว็บเพจภายในเว็บไซต์นั้นเป็นโหนด(node) ภายในทรี ระบบที่ออกแบบได้จัดทำทรีให้กับทุกๆเว็บไซต์เพื่อใช้ในการพิจารณาค่าน้ำหนักรวม โดยแต่ละโหนดของทรีประกอบด้วย ID ที่แสดงถึงแต่ละ URL ที่อยู่บนฐานข้อมูลโดยไม่ซ้ำกัน และค่าน้ำหนักความไม่เหมาะสมของเอกสารเว็บนั้น ตัวอย่างดังรูปที่ 3.9



รูปที่ 3.9 ความสัมพันธ์ของเว็บไซต์และเอกสารเว็บในรูปของ Tree

ค่าน้ำหนักความไม่เหมาะสมของเอกสารเว็บ พิจารณาจาก 2 ปัจจัยคือ ค่าน้ำหนักเฉลี่ยของรูปภาพทั้งหมดบนเอกสารเว็บที่ถูกพิจารณาโดยโปรแกรมตรวจสอบภาพอนาจาร และ ค่าน้ำหนักของข้อความและคุณลักษณะที่ถูกพิจารณาโดยโปรแกรมการตรวจสอบเอกสารเว็บด้วยคำสำคัญ การคำนวณหาค่าน้ำหนักความไม่เหมาะสมนั้นสามารถหาได้จากสมการดังนี้

$$Weight\_self = \frac{(7 * Weight\_image\_avg) + (3 * Weight\_text)}{10} \qquad (3.1)$$

โดย Weight_self คือ ค่าน้ำหนักความไม่เหมาะสมของเอกสารเว็บ , Weight_image_avg คือ ค่าเฉลี่ยน้ำหนักของรูปภาพทั้งหมดบนเอกสารเว็บที่ถูกพิจารณาโดยโปรแกรมตรวจสอบภาพอนาจารหาได้จากสมการที่ 3.2 และ Wight_text คือ ค่าน้ำหนักของข้อความและคุณลักษณะที่ถูกพิจารณาโดยโปรแกรมการตรวจสอบเอกสารเว็บด้วยคำสำคัญ

$$Weight\_image\_avg = \frac{\sum_{i=1}^{n} Weight\_image(i)}{n} \qquad (3.2)$$

n คือจำนวนภาพที่ถูกพิจารณาในหน้าเว็บเพจนั้น และ Weight_image(i) เป็นค่าน้ำหนักของรูปภาพที่ i บนเอกสารเว็บที่ถูกพิจารณาโดยโปรแกรมตรวจสอบภาพอนาจาร

เมื่อคำนวณหาค่าทั้ง 2 ปัจจัยได้แล้วจึงสรุปผลเป็นค่าน้ำหนักความไม่เหมาะสมของเอกสารเว็บตามสมการ 3.1 กำหนดให้ Weight_Image_avg นั้นมีค่าความสำคัญมากกว่า Weight_Text เนื่องจากผลการทดลอง การตรวจสอบด้วยภาพมีความถูกต้องและแม่นยำกว่าการตรวจสอบด้วยคำสำคัญ จึงให้ค่าน้ำหนัก Weight_Image_avg มากกว่า

ส่วนสุดท้ายในการคำนวณค่าน้ำหนักความไม่เหมาะสม คือการสรุปคำนวณค่าน้ำหนักรวมของเว็บไซต์ ซึ่งประกอบด้วยหลายๆ เอกสารเว็บ จากสมการ 3.1 และ 3.2 เป็นเพียงการคำนวณเพื่อหาค่าน้ำหนักความไม่เหมาะสมของเอกสารเว็บเดียว จากที่ได้กล่าวมาแล้วนั้น โครงสร้างของเว็บไซต์สามารถมองในลักษณะของทรีได้ จึงสามารถเขียนสมการความสัมพันธ์ในรูปของความสัมพันธ์โหนดพ่อ (parent node) และ โหนดลูก (child node) เพื่อคำนวณค่าน้ำหนักรวมได้ดังสมการต่อไปนี้

$$Weight\_total = \frac{2 * Weight\_parent + \left(\frac{\sum_{i=1}^{n} Weight\_child(i)}{n} \right)}{3} \qquad (3.3)$$

โดย Weight_total คือ ค่าน้ำหนักความไม่เหมาะสมรวมที่โหนดพ่อ , Weight_parent คือ ค่าน้ำหนักความไม่เหมาะสมของโหนดพ่อ Weight_child คือ ค่าน้ำหนักความไม่เหมาะสมของโหนดลูกซึ่งสามารถคำนวณหา Weight_parent และ Weight_child ได้จากสมการ 3.1 และ 3.2

และ n คือจำนวนโหนดลูกภายใต้โหนดพ่อ กำหนดให้ค่าน้ำหนักของโหนดพ่อมีค่าความสำคัญมากกว่าโหนดลูก ตัวอย่างการคำนวณหาค่าน้ำหนักความไม่เหมาะสม เช่น จากรูปที่ 3.8 กำหนดให้หาค่าน้ำหนักรวมความไม่เหมาะสมที่โหนด W1 โดยมีค่าน้ำหนักที่คำนวณจากสมการ 3.1 และ 3.2 คือ 50 และกำหนดให้ โหนดลูกคือ W5 ,W6 และ W7 มีค่าน้ำหนักที่คำนวณจากสมการ 3.1 และ 3.2 คือ 20 , 30 และ 40 ตามลำดับ จากสมการ 3.3 เราสามารถคำนวณค่าน้ำหนักรวมความไม่เหมาะสมที่โหนด W1 ได้คือ

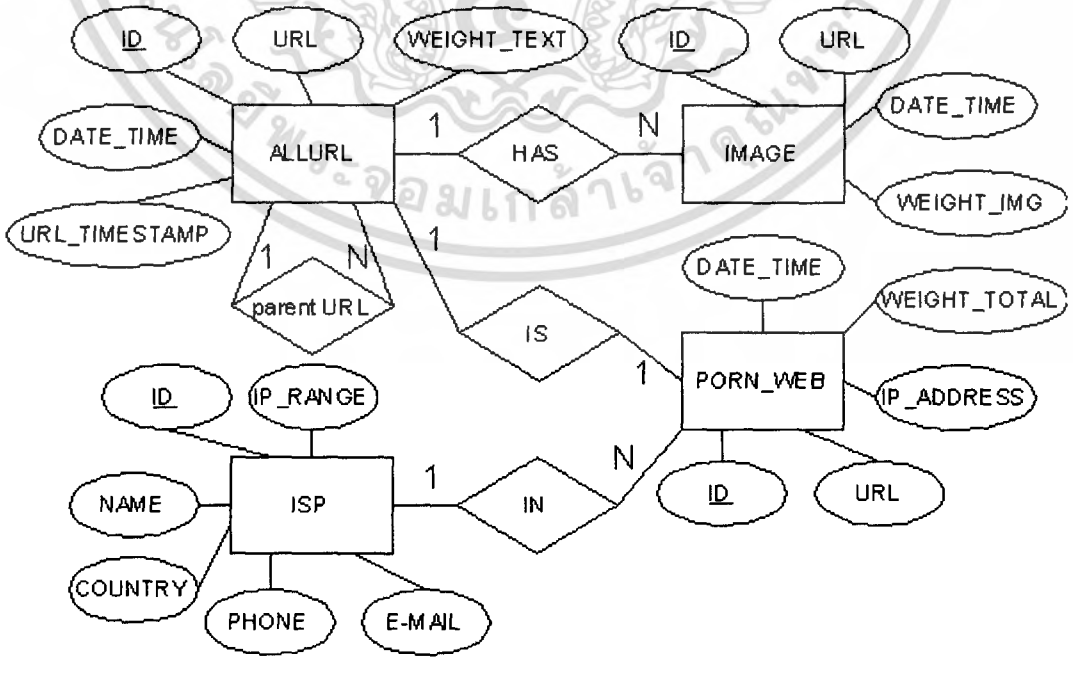
$$Weight\_total = \frac{2*50+\left(\frac{20+30+40}{3}\right)}{3} = 65$$

กระบวนการคำนวณค่าน้ำหนักความไม่เหมาะสมจะดำเนินไปจนกว่าจะสามารถสรุปค่าน้ำหนักรวมของเว็บไซต์ซึ่งเป็นรูทโหนดได้

3.6 การออกแบบฐานข้อมูล

ฐานข้อมูลที่ใช้เก็บข้อมูลของระบบตรวจหาข้อมูลอันตรายในเวิร์ลไวด์เว็บได้ทำการออกแบบดังรูปที่ 3.10 ซึ่งประกอบไปด้วย

1. ตารางเก็บลิงค์ของเว็บเพจทั้งหมดที่โปรแกรมรวบรวมเอกสารเว็บค้นหาได้
2. ตารางเก็บลิงค์ของภาพทั้งหมดที่โปรแกรมรวบรวมเอกสารเว็บค้นหาได้
3. ตารางเก็บลิงค์ของเว็บไซต์ที่ถูกจำแนกว่าเป็นเว็บไซต์อนาจารโดยการผ่านกระบวนการทั้งหมดของโปรแกรมแล้ว
4. ตารางเก็บรายละเอียดของ ISP ที่ดูแลเว็บไซต์อนาจารนั้นๆ



รูปที่ 3.10 ER diagram ของฐานข้อมูลระบบ

ลิงค์ของเว็บเพจทั้งหมดที่โปรแกรมรวบรวมเอกสารเว็บค้นหาได้ จะถูกเก็บในตารางที่ชื่อว่า ALLURL โดยประกอบด้วยฟิลด์ต่างๆ ดังนี้

1. ID ซึ่งเป็นไอดีของลิงค์ของเว็บเพจที่หามาได้ โดยในที่นี้จะกำหนดให้เป็น AUTO_INCREMENT และ กำหนดให้เป็น PRIMARY KEY
2. URL ซึ่งเป็นฟิลด์ที่เก็บลิงค์ของเว็บไซต์ที่โปรแกรมหามาได้ โดยกำหนดให้เป็น NOT NULL และ UNIQUE เพื่อให้ไม่เกิดการซ้ำกันของ URL
3. DATE_TIME ซึ่งเป็นฟิลด์ที่เก็บวันและเวลาที่ค้นพบลิงค์ของเว็บไซต์นั้นๆ
4. URL_TIMESTAMP ซึ่งเป็นฟิลด์ที่เก็บ timestamp ของเว็บเพจนั้นๆ
5. WEIGHT_TEXT ซึ่งเป็นฟิลด์ที่เก็บค่าน้ำหนักของเว็บเพจนั้นๆที่ได้จากค่าน้ำหนักของเนื้อหาและคุณลักษณะที่ถูกพิจารณาโดยโปรแกรมการตรวสอบเนื้อหาของเว็บไซต์
6. URL_ID ซึ่งเป็นฟิลด์ที่เก็บไอดีของเว็บไซต์ตั้งต้นของเว็บเพจนั้นๆ โดยกำหนดเป็น FOREIGN KEY ที่อ้างอิงจาก PRIMARY KEY (ID) ของตาราง ALLURL

ลิงค์ของภาพทั้งหมดที่โปรแกรมรวบรวมเอกสารเว็บค้นหาได้ จะถูกเก็บอยู่ในตารางที่ชื่อว่า IMAGE โดยประกอบด้วยฟิลด์ต่างๆ ดังนี้

1. ID ซึ่งเป็นไอดีของลิงค์ของเว็บไซต์ที่โปรแกรมหามาได้ โดยในที่นี้จะกำหนดให้เป็น AUTO_INCREMENT และ เป็น PRIMARY KEY
2. URL ซึ่งเป็นฟิลด์ที่เก็บลิงค์ของภาพที่โปรแกรมหามาได้ โดยกำหนดค่าให้เป็น NOT NULL และ UNIQUE
3. DATE_TIME ซึ่งเป็นฟิลด์ที่เก็บวันและเวลาที่ค้นพบรูปภาพนั้น
4. WEIGHT_IMG ซึ่งเป็นฟิลด์ที่เก็บค่าน้ำหนักของรูปภาพบนเอกสารเว็บที่ถูกพิจารณาโดยโปรแกรมตรวจสอบภาพอนาจาร

ลิงค์ของเว็บไซต์อนาจารที่ได้จากการผ่านกระบวนการทั้งหมดของโปรแกรมแล้วจะถูกเก็บอยู่ในตารางชื่อ PORN_WEB โดยประกอบด้วยฟิลด์ต่างๆ ดังนี้

1. ID ซึ่งเป็นไอดีของลิงค์เว็บไซต์ทั้งหมดที่ถูกตรวจสอบว่าเป็นเว็บอนาจาร โดยในที่นี้จะกำหนดให้เป็น AUTO_INCREMENT และ เป็น PRIMARY KEY
2. URL ซึ่งเป็นฟิลด์ที่เก็บลิงค์ของเว็บไซต์อนาจาร และ UNIQUE
3. DATE_TIME ซึ่งเป็นฟิลด์ที่เก็บวันและเวลาที่ทำการสรุปว่าเว็บไซต์นั้นเป็นเว็บไซต์อนาจาร
4. IP_ADDRESS ซึ่งเป็นฟิลด์ที่เก็บ IP Address ของ เว็บไซต์นั้นๆ

5. WEIGHT_TOTAL ซึ่งเป็นฟิลด์ที่เก็บค่าน้ำหนักความอันตรายของเว็บไซต์นั้นๆ พิจารณาจาก 2 ปัจจัยคือ ค่าน้ำหนักของรูปภาพทั้งหมดบนเอกสารเว็บที่ถูกพิจารณาโดยโปรแกรมตรวจสอบภาพอนาจาร และ ค่าน้ำหนักของเนื้อหาและคุณลักษณะที่ถูกพิจารณาโดยโปรแกรมการตรวจสอบเนื้อหาของเว็บไซต์

6. ISP_ID ซึ่งเป็นไอดีของ ISP โดยกำหนดเป็น FOREIGN KEY ที่อ้างอิงจาก PRIMARY KEY (ID) ของตาราง ISP

รายละเอียดของ ISP ที่ดูแลเว็บไซต์อนาจารนั้นๆ จะถูกเก็บอยู่ในตารางชื่อ ISP โดยประกอบด้วยฟิลด์ต่างๆ ดังนี้

1. ID ซึ่งเป็นไอดีของ ISP โดยในที่นี้จะกำหนดให้เป็น AUTO_INCREMENT และเป็น PRIMARY KEY
2. NAME ซึ่งเป็นฟิลด์ที่เก็บชื่อของ ISP
3. COUNTRY ซึ่งเป็นฟิลด์ที่เก็บประเทศของ ISP
4. PHONE ซึ่งเป็นฟิลด์ที่เก็บเบอร์โทรศัพท์ติดต่อของ ISP
5. E-MAIL ซึ่งเป็นฟิลด์ที่เก็บอีเมลของ ISP
6. IP_RANGE ซึ่งเป็นฟิลด์ที่เก็บช่วงของไอพีที่ ISP ดูแลอยู่

บทที่ 4

การทดลองและผลการทดลอง

การพัฒนาโปรแกรมแบ่งเป็น 2 ส่วนสำคัญได้แก่ โปรแกรมรวบรวมเอกสารเว็บและโปรแกรมตรวจสอบภาพอนาจาร โดยได้ทดลองเขียนฟังก์ชันการทำงานตามอัลกอริทึมที่ได้ออกแบบไว้ โดยอาศัยโปรแกรมภาษาไพธอน ซึ่งเป็นภาษาระดับสูงและสามารถพัฒนาให้มีประสิทธิภาพ ภาษาไพธอนนั้นมีความยืดหยุ่นสูงสามารถศึกษาและพัฒนาได้โดยง่าย

โปรแกรมรวบรวมเอกสารเว็บจำเป็นต้องทำการติดต่อกับฐานข้อมูลและเครือข่ายภายนอก ในการทดลองจะทำการติดต่อกับฐานข้อมูล MySQL และในส่วนของโปรแกรมตรวจสอบภาพอนาจารนั้นในการทดลองสามารถจำแนกประเภทของภาพได้ด้วยการพัฒนาด้วยภาษาไพธอน

4.1 การทดสอบโปรแกรมรวบรวมเอกสารเว็บ

โปรแกรมรวบรวมเอกสารเว็บจะทำการดึงลิงค์ของหน้าเว็บไซต์นั้นๆมาเก็บไว้ในฐานข้อมูล

<u>วิธีการทดลอง</u>

- ทำการสั่งให้โปรแกรมในส่วนของ Google SOAP Search API ทำงาน เพื่อหาเว็บไซต์ตั้งต้นที่จะให้โปรแกรมรวบรวมเอกสารเว็บทำงานต่อไป
- โปรแกรมรวบรวมเอกสารเว็บจะทำการดึงลิงค์ภายในเว็บไซต์ตั้งต้นแล้วเก็บลงในฐานข้อมูล
- กำหนดให้ฐานข้อมูล MySQL รองรับการเชื่อมต่อได้สูงสุด 500 การเชื่อมต่อและกำหนดให้มี thread ที่ทำงานอยู่ในขณะใดขณะหนึ่งเท่ากับ 128 thread แต่กำหนดให้มี thread ที่ทำการเชื่อมต่อกับฐานข้อมูลได้เพียง 96 thread

<u>ผลการทดลอง</u>

1. โปรแกรมส่วนของ Web Crawler สามารถหาลิงค์ของเว็บเพจได้เฉลี่ย 1093 ลิงค์ต่อนาที สามารถตรวจสอบลิงค์ได้ 148 ลิงค์ต่อนาที ซึ่งเป็นลิงค์ที่สามารถเข้าถึงได้ 115 ลิงค์พบว่าเป็น ลิงค์ที่มีความเสี่ยง 63 ลิงค์ต่อนาที

2. โปรแกรมส่วนของ Web Crawler สามารถหาลิงค์ของรูปภาพได้เฉลี่ย 687 ลิงค์ต่อนาที สามารถตรวจสอบภาพได้ 24 ลิงค์ต่อนาที ซึ่งเป็นลิงค์ที่สามารถเข้าถึงได้และไม่ใช่ภาพขนาดเล็กกว่ากำหนด 17 ลิงค์พบว่าเป็น ภาพที่มีความเสี่ยง 7 ลิงค์ต่อนาที

3. โปรแกรมส่วนของ Web Crawler สามารถหาลิงค์ของเว็บไซต์ได้เฉลี่ย 59เว็บไซต์ต่อนาที สามารถตรวจสอบเว็บไซต์ได้ 7 เว็บไซต์ต่อนาที ซึ่งเป็นเว็บไซต์ที่สามารถเข้าถึงได้ 6 เว็บไซต์พบว่าเป็นเว็บไซต์ที่มีความเสี่ยง 5 เว็บไซต์ต่อนาที

4.2 การทดสอบโปรแกรมตรวจสอบภาพอนาจาร

หลักการตรวจสอบภาพอนาจารที่ประกอบด้วยขั้นตอนการตรวจหาสีผิวมนุษย์ การตรวจสอบคุณสมบัติของภาพและการจัดประเภทของภาพด้วยคุณสมบัติ ซึ่งใช้หลักการของการประมวลผลภาพเข้ามาช่วย เพื่อให้สามารถพิจารณาคุณสมบัติของภาพตามขั้นตอนการตรวจสอบภาพอนาจาร

โครงงานนี้ได้ใช้ภาษาไพธอนในการพัฒนาส่วนของโปรแกรมตรวจสอบภาพอนาจรตามขั้นตอนที่ได้กล่าวไว้ และทดสอบการทำงานของโปรแกรมที่ได้พัฒนาขึ้น ดังนี้

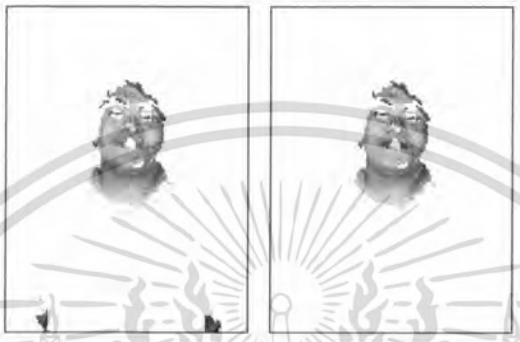
4.2.1 การตรวจหาสีผิวมนุษย์

โครงงานนี้ทดสอบการตรวจหาสีผิวของมนุษย์โดยการพิจารณาพื้นที่ที่มีค่าของพิกเซลอยู่ในช่วงของสีผิวในภาพ ประกอบกับการใช้การตรวจหาขอบของภาพเข้าร่วมการพิจารณาด้วย ซึ่งผลการทดลองเป็นดังรูปที่ 4.1

<u>วิธีการทดลอง</u>

- นำภาพต้นฉบับมาแปลงเป็นภาพในรูปแบบสี RGB ดังรูปที่ 4.1 (a)
- แปลงข้อมูลภาพให้อยู่ในรูปแบบสี YCbCr แล้วผ่านการตรวสอบช่วงของสีผิว ถ้าพิกเซลใดมีค่าไม่อยู่ในช่วงของสีผิวจะถูกแปลงให้เป็นสีขาว ดังรูปที่ 4.1 (b)
- ทำการสร้าง Sobel Edge Operator จากภาพต้นฉบับเพื่อตรวจสอบหาขอบภายในภาพ ดังรูปที่ 4.1(c)
- นำภาพที่ผ่านการตรวจสอบช่วงสีผิวมาตรวจสอบหาขอบภายในภาพ พิกเซลใดที่มีสีอยู่ในช่วงสีผิวแต่อยู่บนขอบจะถูกตัดทิ้ง เพื่อทำการกรองส่วนที่เป็นผิวหนังจริง ดังรูปที่ 4.1 (d)
- ทำการตัดพื้นที่ผิวที่มีขนาดเล็กกว่าขนาดที่กำหนดไว้ออก เพื่อทำการตรวจสอบหาพื้นที่ของสีผิวที่มีขนาดใหญ่ในภาพ ดังรูปที่ 4.1 (e)

(a) (b) (c)



(d) (e)

รูปที่ 4.1(a) ภาพต้นฉบับ

(b) ภาพเมื่อผ่านการทำกระบวนการตัดสีผิว

(c) ภาพเมื่อผ่านการทำ Sobel Edge Oprator กับภาพต้นฉบับ

(d) ภาพ เมื่อผ่านกระบวนการตัดสีผิวและผ่านการกรองด้วย Sobel Edge Oprator

(e) ภาพเมื่อผ่านกระบวนการตัดเอาพื้นที่สีผิวที่มีขนาดเล็กออก

4.2.2 การตรวจสอบคุณสมบัติของภาพ

<u>วิธีการทดลอง</u>

- นำภาพที่ผ่านกระบวนการตรวจหาสีผิวของมนุษย์มาหาพื้นที่ผิวที่ต่อเนื่องกัน
- คำนวณเปอร์เซ็นต์ของจำนวนพิกเซลที่มีสีผิวทั้งหมดในภาพต่อจำนวนพิกเซลทั้งหมดของภาพ
- ทำการพิจารณาหาพื้นที่ผิวที่มีความต่อเนื่องกันซึ่งมีขนาดใหญ่ที่สุดในภาพและคำนวณเปอร์เซ็นต์ของพื้นที่ที่ใหญ่ที่สุดต่อจำนวนพิกเซลทั้งหมดของภาพ
- นับจำนวนพื้นที่ผิวที่ต่อเนื่องกัน
- ทำการพิจารณาพื้นที่ผิวที่มีความต่อเนื่องกันซึ่งมีขนาดใหญ่ที่สุดในภาพ และคำนวณเปอร์เซ็นต์ของความสูงของพื้นที่นั้นต่อความสูงของภาพ
- ทำการพิจารณาพื้นที่ผิวที่มีความต่อเนื่องกันซึ่งมีขนาดใหญ่ที่สุดในภาพ และคำนวณเปอร์เซ็นต์ของความกว้างของพื้นที่นั้นต่อความกว้างของภาพ

- ทำการพิจารณาพื้นที่ผิวที่มีความต่อเนื่องกันซึ่งมีขนาดใหญ่ที่สุดในภาพ และคำนวณหาเปอร์เซ็นต์ของพื้นที่นั้นต่อขนาดพื้นที่ที่มีสีผิวทั้งหมดในภาพ

จากการทดลองการตรวจสอบคุณสมบัติของภาพทั้ง 6 ข้อ ดังที่กล่าวในบทที่ 3 ผลที่ได้จากการทดสอบกับโปรแกรมที่พัฒนาขึ้นด้วยภาษาไพธอน ได้ผลการทดลองดังตารางที่ 4.1

คุณสมบัติของภาพ	ภาพอนาจาร	ภาพไม่อนาจาร
เปอร์เซ็นต์ของสีผิวในภาพ	62.639	6.6233
เปอร์เซ็นต์ของพื้นที่สีผิวที่มีขนาดใหญ่ที่สุด	60.370	5.0526
จำนวนชิ้นส่วนของพื้นที่ผิว	1.8660	1.462
อัตราความสูงของพื้นที่ผิว	93.697	18.9669
อัตราความกว้างของพื้นที่ผิว	93.869	23.8820
อัตราส่วนของพื้นที่ผิวขนาดใหญ่เทียบกับจำนวนพื้นที่ผิวทั้งหมด	93.898	36.33549

ตารางที่ 4.1 ผลการทดลองการตรวจสอบคุณสมบัติของภาพ

จากตารางที่ 4.1 เป็นผลการทดลองโดยใช้ตัวอย่างภาพ 1000 ภาพที่มีคุณสมบัติเป็นภาพอนาจารและ ภาพ 1000 ภาพเป็นภาพไม่อนาจาร ทำการคำนวณหาค่าของคุณสมบัติของภาพ แล้วทำการหาค่าเฉลี่ย จากการทดลองจะสังเกตเห็นว่าภาพอนาจารจะประกอบไปด้วยพื้นที่ของสีผิวขนาดใหญ่และครอบคลุมพื้นที่ส่วนใหญ่ของภาพ

4.2.3 การจัดประเภทของภาพด้วยคุณสมบัติ

จากการทดลองการตรวจสอบคุณสมบัติของภาพจะเห็นว่าบางคุณสมบัติไม่สามารถใช้เป็นตัวระบุได้ว่าภาพนั้นเป็นภาพอนาจารได้อย่างชัดเจน ดังนั้นจึงต้องมีการนำคุณสมบัติของภาพทั้ง 6 อย่างมาผ่านหลักการการวิเคราะห์ส่วนประกอบ (Principle Component Analysis) โดยนำคุณสมบัติทั้ง 6 มาพิจารณาร่วมกันเพื่อให้สามารถระบุว่าภาพนั้นเป็นภาพอนาจารได้อย่างถูกต้อง

<u>วิธีการทดลอง</u>

- นำตัวอย่างภาพอนาจาร 1,000 ภาพ และภาพที่ไม่ใช่ภาพอนาจารที่ประกอบด้วยภาพมนุษย์, ภาพสัตว์ , พืช , เมือง , ทิวทัศน์ และอื่นๆ 1,000 ภาพ มาผ่านกระบวนการตรวจหาสีผิวมนุษย์และการตรวจสอบคุณสมบัติของภาพ
- รวบรวมคุณสมบัติทั้ง 6 ข้อของภาพ 2,000 ภาพมาผ่านกระบวนการวิเคราะห์ส่วนประกอบ จากทฤษฎีในหัวข้อ 2.9.3 และการออกแบบพัฒนาในหัวข้อ 3.3.4 เรื่องกระบวนการวิเคราะห์ส่วนประกอบ จำเป็นต้องทำการคำนวณหาค่าของโควาเรียนซ์

เมตริกซ์ ของชุดข้อมูลที่ทำการปรับค่าแล้ว (Data Adjust) จากการทดลองกับข้อมูลตัวอย่าง 2000 ภาพซึ่งเป็นภาพอนาจาร 1000 ภาพ และ ภาพไม่อนาจาร 1000 ภาพ สามารถคำนวณหาค่าของโควาเรียนซ์เมตริกซ์ ขนาด 6*6 (จากคุณสมบัติของภาพ 6 ประการ) ของชุดข้อมูลที่ทำการปรับค่าแล้ว (Data Adjust) จากสมการ 2.19 ได้ดังนี้

[[1.07365201e+03,1.06818650e+03,4.00028861e+01,1.28221037e+03,1.24811587e+03,1.01658911e+03]
[1.06818650e+03,1.07955006e+03,1.69821372e+00,1.26294069e+03,1.22216130e+03,1.01956179e+03]
[4.00028861e+01,1.69821372e+00,1.84891499e+02,1.38354192e+02,1.68839558e+02,1.24626338e+02]
[1.28221037e+03,1.26294069e+03,1.38354192e+02,1.84434595e+03,1.71466417e+03,1.53120164e+03]
[1.24811587e+03,1.22216130e+03,1.68839558e+02,1.71466417e+03,1.86008204e+03,1.55952563e+03]
[1.01658911e+03,1.01956179e+03,1.24626338e+02,1.53120164e+03,1.55952563e+03,1.74865527e+03]]

ขั้นตอนต่อไปต้องทำการคำนวณไอแกนแวลูส์ และ ไอแกนเวกเตอร์ ของโควาเรียนซ์เมตริกซ์ ซึ่งเป็นเมตริกซ์ขนาด 6*6 เราสามารถคำนวณค่าของไอแกนแวลูส์ และไอแกนเวกเตอร์ได้ 6 คู่ดังนี้

ค่าไอแกนแวลูส์ คำนวณจากสมการ 2.22

[6.81879802e+03,5.16192545e+02,3.74782796e+00,2.40435040e+02,8.27591828e+01,1.29244210e+02]

ค่าไอแกนเวกเตอร์ จากการแทนค่าไอแกนแวลูส์ที่คำนวณได้ในสมการ 2.21

[[0.37311171 , 0.37007975 , 0.03439508 , 0.50656072 , 0.50569109 , 0.45860481]
[-0.46076256 ,-0.48284077 , 0.2327821 ,-0.05986624 , 0.10544402 , 0.69690225]
[-0.7093189 , 0.68922908 , 0.13863331 , 0.02099868 , 0.02444884 ,-0.03964963]
[0.06904139 , 0.25620754 ,-0.67041743 ,-0.20335488 ,-0.43899494 , 0.49604606]
[0.37102894 , 0.28529706 , 0.66889249 ,-0.39878516 ,-0.34377693 , 0.23730523]
[-0.05406512 ,-0.08850044 , 0.16891889 , 0.73415837 ,-0.64935445 , 0.00783144]]

โดยค่าของไอแกนแวลูส์จะเป็นคู่กับไอแกนเวกเตอร์เช่น ค่า 6.81879802e+03 เป็นไอแกนแวลูส์คู่กับไอแกนเวกเตอร์ [0.3731117 , 0.37007975, 0.03439508 , 0.50656072 , 0.50569109 , 0.45860481] เป็นต้น

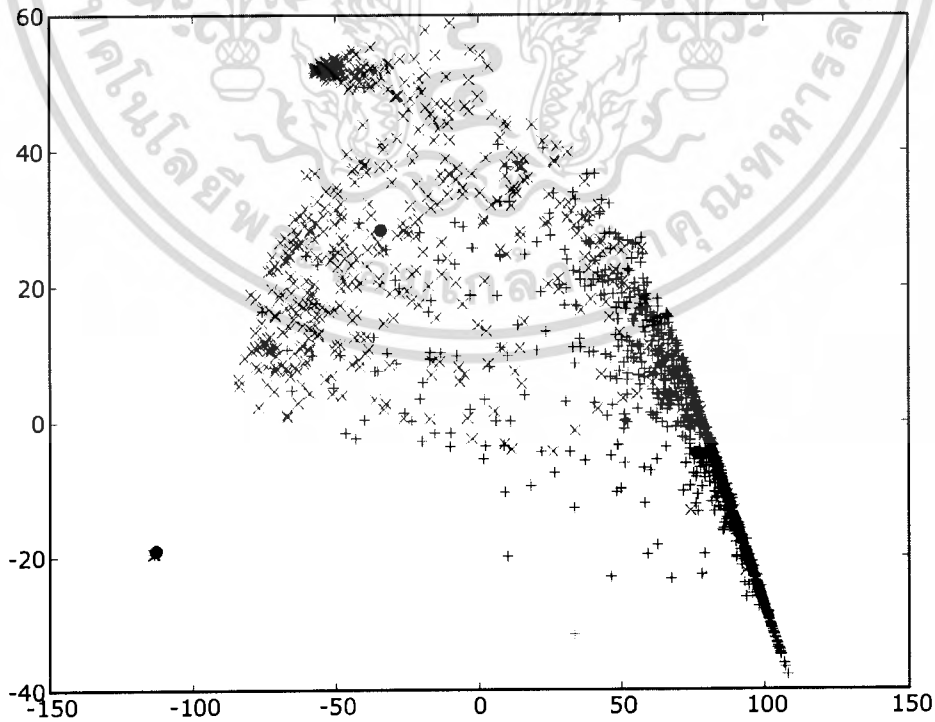
- เลือกคุณสมบัติที่มีความสำคัญสูงที่สุด 2 คุณสมบัติที่ทำให้สูญเสียข้อมูลน้อยที่สุด จากค่าไอแกนแวลูส์ และไอแกนเวกเตอร์ที่คำนวณได้ ทำการเลือกไอแกนเวกเตอร์ที่มีค่าไอแกนแวลูส์สูงที่สุด 2 อันดับแรก คือ 6.81879802e+03 และ 5.16192545e+02 ซึ่งมีไอแกนเวกเตอร์คือ [0.37311171 , 0.37007975 , 0.03439508 , 0.50656072

0.50569109, 0.45860481] และ [-0.46076256 ,-0.48284077 , 0.2327821 ,-0.05986624 , 0.10544402 ,0.69690225] ตามลำดับ

- ทำการหาค่าของชุดข้อมูลสุดท้าย (Final Data) จากสมการ 2.23 ซึ่งต้องใช้ไอแกนเวกเตอร์ที่คำนวณได้ และชุดของข้อมูลที่ปรับค่าแล้ว (Data Adjust) หลังจากนั้นนำชุดข้อมูลสุดท้ายมาผ่านกระบวนการจัดกลุ่ม (Clustering) ด้วยกระบวนการ K-mean เมื่อกำหนดค่า K = 3 เพื่อหาจุดเซ็นทรอยด์แบ่งข้อมูลเป็น 3 กลุ่ม ซึ่งสามารถคำนวณหาจุดเซ็นทรอยด์ ได้คือ จุดพิกัด [76.3960,-4.6239] ซึ่งแสดงถึงข้อมูลที่เป็นภาพอนาจาร ,จุดพิกัด [-34.1640,28.3840] ซึ่งแสดงถึงข้อมูลที่ไม่ใช่ภาพอนาจารแต่มีสีผิวเป็นส่วนประกอบ และ จุดพิกัด [-113.0394,-18.9615] ซึ่งแสดงถึงข้อมูลที่ไม่ใช่ภาพอนาจารและไม่มีสีผิวเป็นส่วนประกอบ
- กระบวนการขั้นต้นเป็นกระบวนการเตรียมค่าสำหรับการจำแนกภาพ เมื่อต้องการจำแนกภาพ สามารถทำได้โดยใช้ K-Nearest Neighbor algorithm กับทุกข้อมูลภาพที่ปรับค่าแล้วเทียบกับจุดเซ็นทรอยด์ทั้ง 3 จุด

<u>การทดลองที่ 1</u>

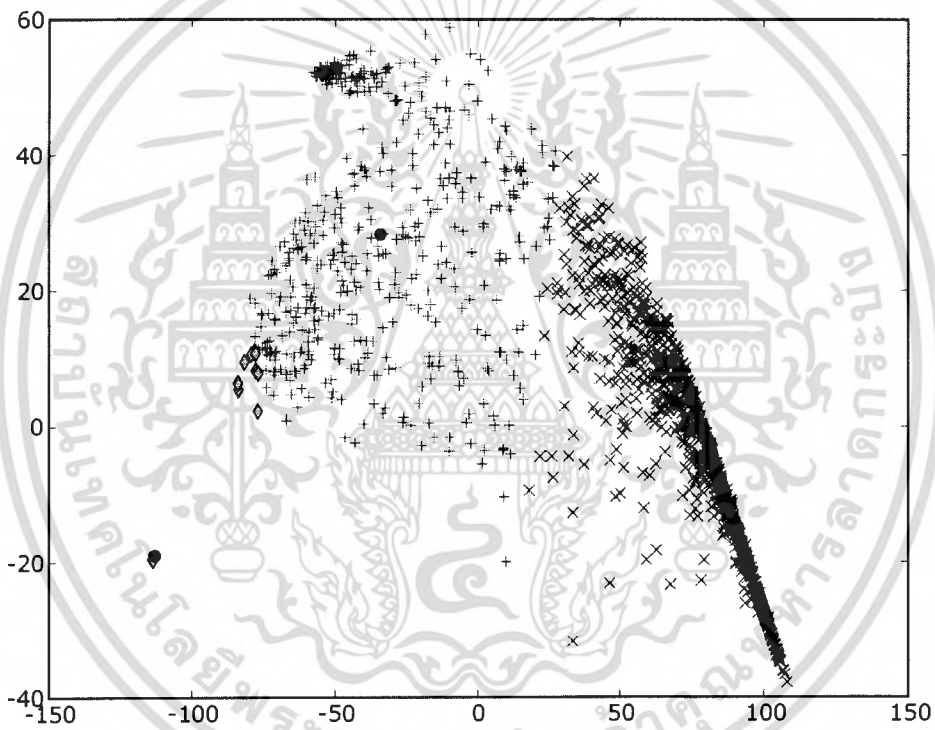
จากการทดลองพัฒนาโปรแกรมการจัดประเภทของภาพด้วยคุณสมบัติด้วยภาษาไพธอนจะได้ผลดังรูปที่ 4.2



รูปที่ 4.2 กราฟของความสัมพันธ์ของคุณสมบัติของภาพ

จากรูปที่ 4.2 แสดงกราฟความสัมพันธ์ของคุณสมบัติของภาพตัวอย่าง 2,000 ภาพ โดยแบ่งออกเป็น 2 กลุ่ม ได้แก่ ภาพอนาจาร, ภาพไม่อนาจาร ซึ่งภาพอนาจารแสดงด้วยสัญลักษณ์ '+' ซึ่งรวมตัวกันอยู่ทางมุมล่างขวาของกราฟ และภาพไม่อนาจารแสดงด้วยสัญลักษณ์ 'x' ซึ่งรวมตัวกันอยู่ตรงกลางและมุมล่างขวาของกราฟ แต่ด้วยกระบวนการของการจัดกลุ่ม K-Mean clustering สามารถแบ่งกลุ่มข้อมูลออกเป็น 3 กลุ่ม ได้แก่ ภาพอนาจาร , ภาพไม่อนาจารแต่มีสีผิวเป็นส่วนประกอบ และภาพที่ไม่มีสีผิวเป็นส่วนประกอบ

การทดสอบการจัดกลุ่มข้อมูลด้วยตัวอย่างภาพที่นำมาวิเคราะห์องค์ประกอบ 2,000 ภาพ ได้ผลการทดลองดังรูปที่ 4.3



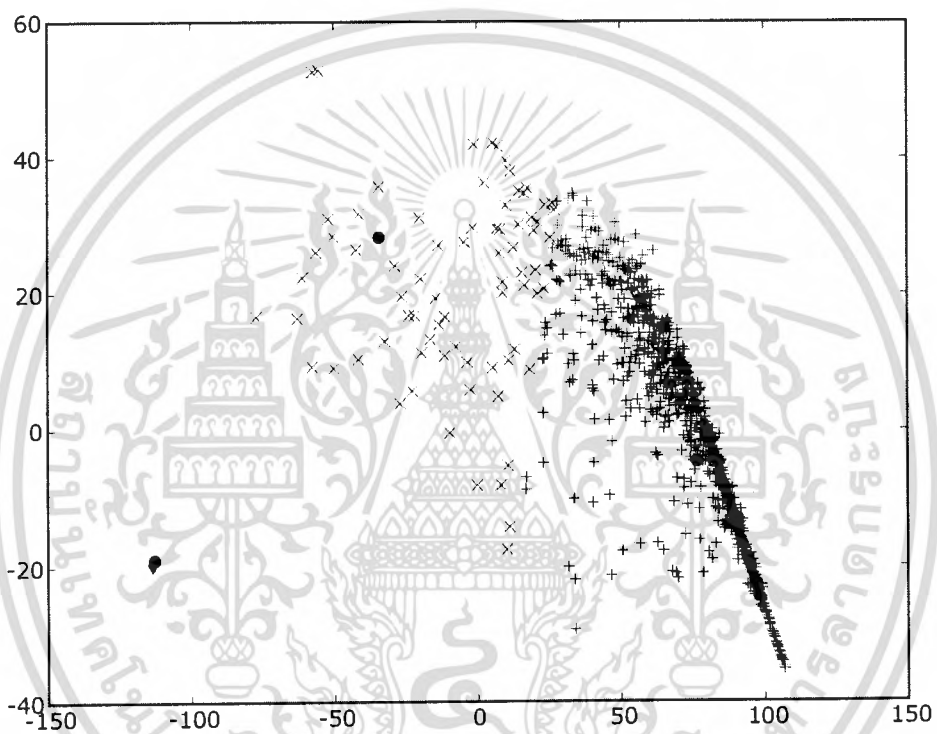
รูปที่ 4.3 ผลการทดลองการจัดกลุ่มข้อมูลของตัวอย่างภาพ

ประเภทของภาพ	จำนวนภาพ	ภาพอนาจาร	ภาพที่ไม่ใช่ภาพอนาจาร	ความถูกต้อง (%)	ความผิดพลาด (%)
ภาพอนาจาร	1000	919	81	91.9	8.1
ภาพที่ไม่ใช่ภาพอนาจาร	1000	65	935	93.5	6.5
รวม	2000	984	1016	92.7	7.3

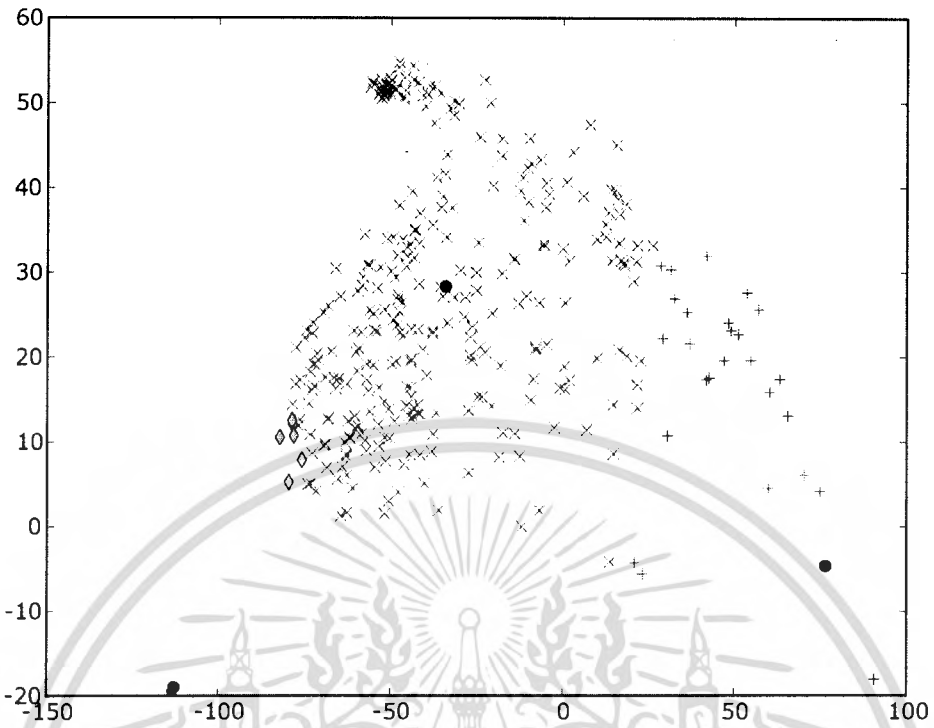
ตารางที่ 4.2 ประสิทธิภาพของการจัดกลุ่มข้อมูลของตัวอย่างภาพ

การจัดกลุ่มข้อมูลมีการแบ่งออกเป็น 3 กลุ่ม ได้แก่ ภาพอนาจาร, ภาพที่ไม่ใช่ภาพอนาจารแต่มีสีผิวเป็นส่วนประกอบ และภาพที่ไม่มีสีผิวเป็นส่วนประกอบ แสดงดังรูปที่ 4.3 โดยภาพอนาจารแสดงแทนด้วยสัญลักษณ์ 'x' , ภาพที่ไม่ใช่ภาพอนาจารแต่มีสีผิวเป็นส่วนประกอบแสดงแทนด้วยสัญลักษณ์ '+' และภาพที่ไม่มีส่วนประกอบของสีผิวแสดงแทนด้วยสัญลักษณ์ '◇ ' และประสิทธิภาพของการทำงานแสดงดังตารางที่ 4.2

การทดลองที่ 2



รูปที่ 4.4 ผลการแบ่งกลุ่มของภาพอนาจารนอกกลุ่มตัวอย่าง



รูปที่ 4.5 ผลการแบ่งกลุ่มของภาพที่ไม่ใช่ภาพอนาจารนอกกลุ่มตัวอย่าง

ประเภทของภาพ	จำนวนภาพ	ภาพอนาจาร	ภาพที่ไม่ใช่ภาพอนาจาร	ความถูกต้อง (%)	ความผิดพลาด (%)
ภาพอนาจาร	1000	924	76	92.4	7.6
ภาพที่ไม่ใช่ภาพอนาจาร	500	26	474	94.8	5.2
รวม	1500	950	550	93.2	6.8

ตารางที่ 4.3 ประสิทธิภาพของการจัดกลุ่มข้อมูลของภาพนอกกลุ่มตัวอย่าง

การทดลองนี้ใช้ภาพที่ไม่ได้อยู่ในกลุ่มภาพตัวอย่างจำนวน 1,500 ภาพ ซึ่งประกอบด้วยภาพอนาจาร 1,000 ภาพ และภาพที่ไม่ใช่ภาพอนาจาร 500 ภาพ ด้วยการจัดกลุ่มโดยใช้จุดเซ็นทรอยด์ที่หาได้จากการทดลองที่ 1 โดยภาพอนาจารได้ผลดังรูปที่ 4.4 และภาพที่ไม่ใช่ภาพอนาจารได้ผลดังรูปที่ 4.5

จากรูปที่ 4.4 และ 4.5 แสดงการแบ่งกลุ่มของภาพอนาจารและภาพที่ไม่ใช่ภาพอนาจาร ภาพที่ถูกตรวจสอบว่าเป็นภาพอนาจารแสดงด้วยสัญลักษณ์ '+' , ภาพที่ถูกตรวจสอบว่าไม่ใช่ภาพอนาจารแต่มีสีผิวเป็นส่วนประกอบแสดงด้วยสัญลักษณ์ 'x' และ ภาพที่ถูกตรวจสอบว่าไม่มี

ส่วนประกอบของสีผิวแสดงแทนด้วยสัญลักษณ์ ' ◇ ' และประสิทธิภาพการทำงานแสดงดังตารางที่ 4.3

ผลการทดสอบประสิทธิภาพ

โปรแกรมส่วนของ Image Detection สามารถตรวจสอบและคำนวณค่าน้ำหนักของรูปได้โดยใช้เวลา เฉลี่ย 1.724 วินาทีต่อภาพ

4.3 การทดสอบโปรแกรมการวิเคราะห์เนื้อหาในเว็บไซต์

การตรวจสอบเนื้อหาของเว็บเพจจะประกอบด้วยขั้นตอนการตรวจสอบคุณสมบัติของเว็บเพจและการจัดประเภทของเนื้อหาเว็บเพจด้วยคุณสมบัติ ซึ่งใช้หลักการ Regular Expression เข้ามาช่วยในการตรวจสอบ ขั้นตอนการทดลองการทำงานของโปรแกรมที่ได้พัฒนาขึ้น มีดังนี้

4.3.1 การตรวจสอบคุณสมบัติของเนื้อหาในเว็บเพจ

วิธีการทดลอง

- ดึงลิงค์เว็บเพจจากฐานข้อมูลที่ได้มาจากโปรแกรมรวบรวมเอกสารเว็บ
- ทำการค้นหาจำนวนของ IMG Tag ในเอกสารเว็บของเว็บเพจนั้น
- ทำการค้นหาจำนวนของ META Tag ในเอกสารเว็บของเว็บเพจนั้น
- ทำการค้นหาจำนวนของคำที่ไม่เหมาะสมใน TITLE Tag และ META Tag ในเอกสารเว็บของเว็บเพจนั้น
- ทำการค้นหาจำนวน SCRIPT Tag ในเอกสารเว็บของเว็บเพจนั้น
- ทำการค้นหาจำนวนลิงค์ทั้งหมดในเอกสารเว็บของเว็บเพจนั้น
- ทำการค้นหาจำนวนลิงค์ที่เป็นรูปภาพในเอกสารเว็บของเว็บเพจนั้น
- ทำการค้นหาจำนวนลิงค์ที่เป็นข้อความในเอกสารเว็บของเว็บเพจนั้น
- ทำการค้นหาจำนวนลิงค์ที่ไม่ใช่ลิงค์ภายในในเอกสารเว็บของเว็บเพจนั้น
- ทำการค้นหาจำนวนลิงค์ภายในในเอกสารเว็บของเว็บเพจนั้น
- ทำการค้นหา PARAM Tag ในเอกสารเว็บของเว็บเพจนั้น

ผลการทดลอง

จากการทดลองการตรวจสอบคุณสมบัติของเนื้อหาในเว็บไซต์ทั้ง 10 ข้อ ดังที่กล่าวในบทที่ 3 ผลที่ได้จากการทดสอบกับโปรแกรมที่พัฒนาขึ้นด้วยภาษาไพธอน ได้ผลการทดลองซึ่งมาจากค่าเฉลี่ยคุณสมบัติทั้ง 10 ประการของเว็บไซต์อนาจาร 100 เว็บไซต์และเว็บไซต์ไม่อนาจาร 100 เว็บไซต์ ดังตารางที่ 4.4

คุณสมบัติของเนื้อหาในเว็บเพจ	เว็บไซต์อนาจาร	เว็บไซต์ไม่อนาจาร
จำนวน IMG Tag	51.39	37.40
จำนวน META Tag	4.80	5.41
จำนวนคำที่ไม่เหมาะสม ใน TITLE Tag และ META Tag	14.30	0.84
จำนวน SCRIPT Tag	2.42	6.86
จำนวนลิงค์ทั้งหมด	290.89	75.80
จำนวนลิงค์ที่เป็นรูปภาพ	32.56	15.14
จำนวนลิงค์ที่เป็นข้อความ	258.33	60.66
จำนวนลิงค์ที่ไม่ใช่ที่ลิงค์ภายใน	126.33	31.63
จำนวนลิงค์ที่เป็นลิงค์ภายใน	164.56	44.17
จำนวน PARAM Tag	0.10	1.27

ตารางที่ 4.4 ผลทดลองการตรวจสอบคุณสมบัติของเนื้อหาในเว็บไซต์

4.3.2 การจัดประเภทของเว็บเพจด้วยคุณสมบัติ

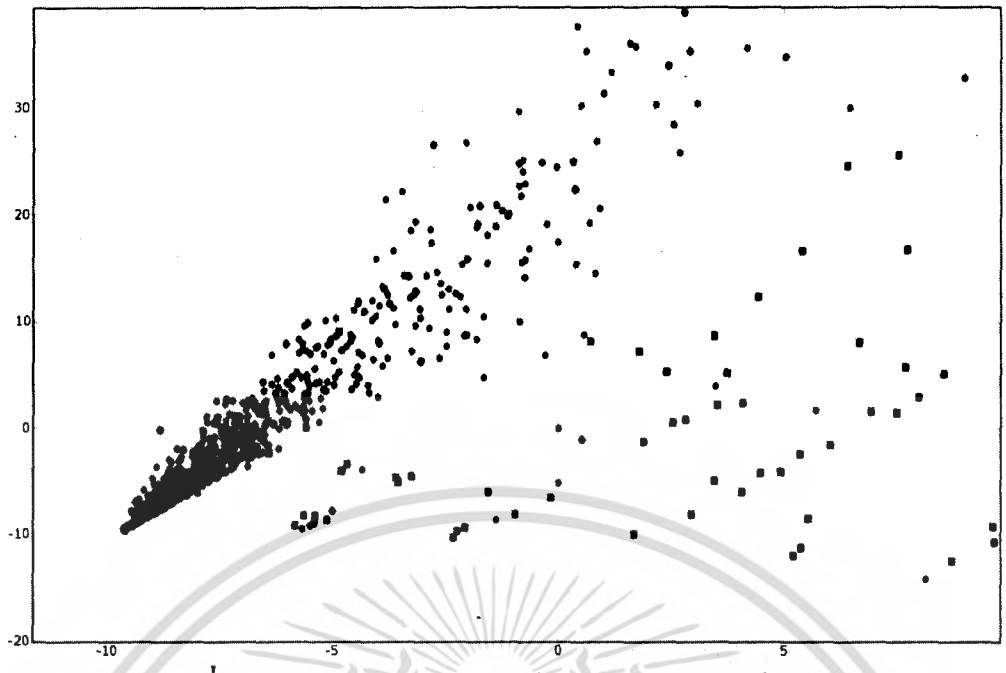
จากการทดลองการตรวจสอบคุณสมบัติของเนื้อหาในเว็บเพจจะเห็นว่าบางคุณสมบัติไม่สามารถใช้เป็นตัวระบุได้ว่าเว็บเพจนั้นเป็นเว็บอนาจารได้อย่างชัดเจน ดังนั้นจึงต้องมีการนำคุณสมบัติของเว็บเพจทั้ง 10 ประการมาผ่านหลักการการวิเคราะห์ส่วนประกอบ (Principle Component Analysis) โดยนำคุณสมบัติทั้ง 10 มาพิจารณาร่วมกันเพื่อให้สามารถระบุว่าเว็บนั้นเป็นเว็บอนาจารได้อย่างถูกต้อง

<u>วิธีการทดลอง</u>

- นำตัวอย่างลิงค์ของเว็บเพจที่เป็นเว็บอนาจารจำนวน 201 ลิงค์ และลิงค์ของเว็บเพจที่ไม่ใช่เว็บอนาจารอีก 961 ลิงค์
- รวบรวมคุณสมบัติทั้ง 10 ประการของเว็บเพจมาผ่านกระบวนการวิเคราะห์ส่วนประกอบ
- เลือกคุณสมบัติที่มีความสำคัญสูงที่สุด 2 คุณสมบัติที่ทำให้สูญเสียข้อมูลน้อยที่สุด
- หาสมการเส้นตรงที่ตัดแบ่งกลุ่มของข้อมูล

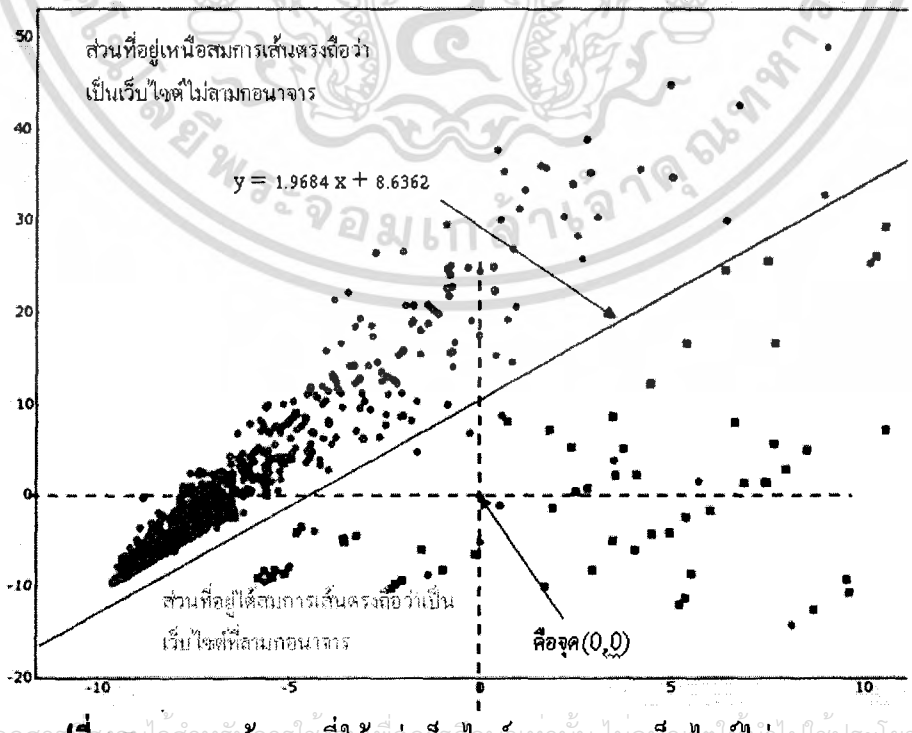
<u>ผลการทดลอง</u>

เมื่อนำคุณสมบัติทั้ง 10 ประการมาผ่านกระบวนการวิเคราะห์ส่วนประกอบและนำมาแสดงให้อยู่ในรูปกราฟจะได้ผลดังรูปที่ 4.6



รูปที่ 4.6 กราฟของความสัมพันธ์ของคุณสมบัติของเว็บเพจ

จากรูปที่ 4.6 แสดงข้อมูลที่เป็นลิงค์ของเว็บไซต์อนาจารจำนวน 201 ลิงค์ และลิงค์ของเว็บเพจที่ไม่ใช่เว็บไซต์อนาจารอีก 961 ลิงค์ ซึ่งเว็บไซต์อนาจารแสดงด้วยสัญลักษณ์วงกลมและเว็บไซต์ไม่อนาจารแสดงด้วยสัญลักษณ์สี่เหลี่ยม และเมื่อพล็อตจุดข้อมูลทั้งหมดของกลุ่มตัวอย่างแล้วจะทำการหาสมการเส้นตรงที่ตัดแบ่งกลุ่มข้อมูลออกจากกัน ซึ่งจุดข้อมูลที่อยู่เหนือเส้นสมการจะถือว่าเป็นเว็บไซต์ไม่อนาจารส่วนข้อมูลที่อยู่ใต้สมการถือว่าเป็นเว็บไซต์อนาจาร ดังแสดงในรูปที่ 4.7



รูปที่ 4.7 สมการเส้นตรงที่ใช้แบ่งเว็บไซต์อนาจารจากเว็บไซต์ไม่อนาจาร

จากสมการเส้นตรง y = mx+b หากมีการกำหนดให้ความชันคงที่และเปลี่ยนแปลงค่า x,y จะได้ค่า b ที่แตกต่างออกไป ซึ่งสามารถรู้ได้ว่าอยู่เหนือสมการเส้นตรงหรืออยู่ใต้สมการเส้นตรง ซึ่งในที่นี้ได้สมการเส้นตรงเป็น y = 1.9684 x + 8.6362 นั่นคือ

- หากค่า b มากกว่า 8.6362 ถือว่าเป็นเว็บไซต์ไม่อนาจาร
- หากค่า b น้อยกว่าหรือเท่ากับ 8.6362 ถือว่าเป็นเว็บไซต์อนาจาร

ประเภทของเว็บ	จำนวนเว็บ	เว็บอนาจาร	เว็บไม่อนาจาร	ความถูกต้อง (%)	ความผิดพลาด (%)
เว็บอนาจาร	201	185	16	91.35	8.65
เว็บไม่อนาจาร	961	945	16	98.33	1.67
รวม	1162	1130	32	97.25	2.75

ตารางที่ 4.5 ประสิทธิภาพของการจัดกลุ่มข้อมูลของเว็บไซต์กลุ่มตัวอย่าง

เมื่อทำการทดสอบโปรแกรมกับข้อมูลตัวอย่างเพื่อหาสมการเส้นตรงที่แบ่งแยกเว็บไซต์อนาจารจากเว็บไซต์ไม่อนาจารได้แล้ว นำข้อมูลนอกกลุ่มตัวอย่างซึ่งประกอบด้วยเว็บไซต์อนาจารจำนวน 300 เว็บไซต์ และเว็บไซต์ไม่อนาจารจำนวน 1224 เว็บไซต์มาทดสอบความถูกต้องซึ่งใช้สมการเส้นตรงเดียวกันในการแบ่งแยกกลุ่ม ได้ผลการทดลองดังแสดงในตารางที่ 4.6

ประเภทของเว็บ	จำนวนเว็บ	เว็บอนาจาร	เว็บไม่อนาจาร	ความถูกต้อง (%)	ความผิดพลาด (%)
เว็บอนาจาร	300	265	35	88.33	11.67
เว็บไม่อนาจาร	1224	1086	138	89.02	10.98
รวม	1524	1351	173	88.65	11.35

ตารางที่ 4.6 ประสิทธิภาพของการจัดกลุ่มข้อมูลของเว็บไซต์นอกกลุ่มตัวอย่าง

ผลการทดสอบประสิทธิภาพ

โปรแกรมส่วนของการตรวจสอบเนื้อหาของเว็บเพจสามารถตรวจสอบคุณสมบัติและคำนวณค่าน้ำหนักของเว็บเพจได้โดยใช้เวลาเฉลี่ย 1.560 วินาทีต่อเว็บเพจ

4.4 การทดสอบประสิทธิภาพระบบ

ผลการทดสอบประสิทธิภาพการใช้ทรัพยากรของระบบ

1. เมื่อโปรเซสทั้งสามทำงานร่วมกันจะใช้ CPU ในการประมวลผลโดยเฉลี่ย 88.7%
2. เมื่อโปรเซสทั้งสามทำงานร่วมกันจะใช้หน่วยความจำโดยเฉลี่ย 75.7 %

3. เมื่อโปรเซสทั้งสามทำงานร่วมกันจะใช้ทรัพยากรเครือข่ายในการรับข้อมูลโดยเฉลี่ย 149.66 KB/s และในการส่งข้อมูลโดยเฉลี่ย 10.42 KB/s

ผลการทดสอบประสิทธิภาพโดยรวม

จากกลุ่มตัวอย่างสามารถคำนวณประสิทธิภาพและความผิดพลาดได้ดังนี้

1. โปรแกรมส่วนของการตรวจสอบรูปภาพมีอัตราความถูกต้อง 92.7% และอัตราความผิดพลาด 7.3%

2. โปรแกรมส่วนของการตรวจสอบเนื้อหาของเว็บเพจมีอัตราความถูกต้อง 97.25% และอัตราความผิดพลาด 2.75%

3. โปรแกรมทั้งหมดมีอัตราความถูกต้อง 97.6% และอัตราความผิดพลาด 2.4%

บทที่ 5
บทวิจารณ์และสรุป

5.1 บทสรุป

โครงงานนี้เป็นโครงงานที่พัฒนาโปรแกรมการตรวจหาข้อมูลอันตรายในเวิร์ลไวด์เว็บ ซึ่งในโครงงานจะมีการทำงานแบ่งเป็น 3 ส่วนหลักๆ คือ

1. ส่วนของโปรแกรมรวบรวมเอกสารเว็บ เป็นโปรแกรมที่ทำการดึงลิงค์ของแต่ละหน้าเว็บเพจเก็บลงในฐานข้อมูล ซึ่งจากการทดลองมีการทำงานเป็นแบบมัลติเทรดที่จำกัดจำนวนของการเชื่อมต่อ ทำให้การทำงานของโปรแกรมรวบรวมเอกสารเว็บรวดเร็วและมีประสิทธิภาพมากขึ้น แต่อาจมีความผิดพลาดเนื่องจากกรณี เช่น การร้องขอเอกสารเว็บจากเครื่องที่ให้บริการบนเครือข่ายที่มีความล่าช้าทำให้ time out ก่อนได้รับเอกสาร หรือ เอกสารที่ได้รับมาไม่สามารถค้นหาข้อมูลที่ต้องการภายในเอกสารนั้นได้ ได้แก่ เอกสารที่ไม่อยู่ในรูปแบบของ HTML

2. ส่วนของโปรแกรมการตรวจสอบภาพอนาจาร ซึ่งเป็นส่วนที่นำภาพที่โปรแกรมรวบรวมเอกสารเว็บหามาได้นำมาผ่านกระบวนการประมวลผลภาพเพื่อตรวจสอบว่าภาพนั้นๆ เป็นภาพอนาจารหรือไม่ จากการทดลองยังมีความผิดพลาดอยู่ ดังนี้

- กระบวนการตัดสีผิว มีความผิดพลาดเกิดขึ้นเนื่องจาก
  - ภาพมีคุณภาพต่ำ
  - ภาพมีวัตถุที่มีสีใกล้เคียงกับสีผิวมนุษย์ เช่น ภาพของพืชที่มีสีเหลือง , ภาพคนที่ใส่เสื้อที่มีสีใกล้เคียงกับสีผิว , ภาพใบหน้าคนขนาดใหญ่
  - ภาพผิวหนังที่มีสีผิวจางเนื่องจากการส่องสว่างและการสะท้อนของแสง
  - ภาพขนาดใหญ่ที่ประกอบขึ้นจากภาพขนาดเล็กหลายภาพ
- กระบวนการวิเคราะห์องค์ประกอบ มีความผิดพลาดเกิดขึ้นเนื่องจากการรวมคุณสมบัติ และตัดคุณสมบัติบางส่วนที่มีความสำคัญต่ำ ทำให้สูญเสียข้อมูลบางส่วนไป อย่างไรก็ตามประสิทธิภาพการทำงานของโปรแกรมตรวจสอบภาพอนาจารเป็นที่น่าพอใจ สามารถจำแนกความแตกต่างระหว่างภาพอนาจารและภาพที่ไม่ใช่ภาพอนาจารได้

3. ส่วนของโปรแกรมการตรวจสอบเนื้อหาในเว็บไซต์ ซึ่งเป็นส่วนที่จะพิจารณาเนื้อหาในหน้าเว็บเพจนั้นจะใช้กระบวนการดึงส่วนที่เป็นเอกสารเว็บของเว็บเพจมาทำการตรวจสอบเพื่อพิจารณาคุณสมบัติต่างๆภายในหน้าเว็บเพจนั้นๆและนำมาหาค่าความอันตรายของแต่ละเว็บเพจ ซึ่งอาจมีความผิดพลาดได้เนื่องจากเว็บไซต์ไม่อนาจารบางเว็บไซต์มีคุณสมบัติของเนื้อหาในเว็บไซต์ใกล้เคียงกับเว็บไซต์อนาจาร

เมื่อผ่านกระบวนการ 3 ส่วนหลักๆดังที่กล่าวมาแล้วจะสามารถพิจารณาได้ว่าเว็บไซต์ดังกล่าวเป็นเว็บไซต์อนาจารหรือไม่ โดยคิดเป็นค่าน้ำหนักความอันตรายของเว็บไซต์นั้นๆ และ

จะเก็บรายชื่อเว็บไซต์ไว้ในฐานข้อมูลเพื่อให้โปรแกรมสามารถดึงข้อมูลเพื่อนำไปแสดงบนหน้าเว็บไซต์ที่จัดทำขึ้น

5.2 วิจารณ์สิ่งที่ได้จากโครงงาน

สิ่งที่ได้จากโครงงานนั้นคือทำให้ผู้พัฒนาได้มีความรู้เกี่ยวกับ กระบวนการทำงานของโปรแกรมรวบรวมเอกสารเว็บ การตรวจสอบภาพอนาจารด้วยวิธีการของการประมวลผลภาพและการจำแนกภาพด้วยคุณสมบัติ พร้อมทั้งกระบวนการในการวิเคราะห์ลักษณะจำเพาะของเว็บไซต์อนาจารอีกด้วย สำหรับผู้ที่นำโครงงานนี้ไปใช้งาน ผู้พัฒนาหวังว่าจะได้รับความรู้เพียงพอที่จะสามารถเข้าใจถึงหลักการทำงาน และสามารถประยุกต์การใช้งานระบบได้ จากจุดประสงค์ของโครงงานที่ต้องการจัดทำรายการเว็บไซต์ต้องห้ามเพื่อให้ระบุตัวตนของเว็บไซต์เหล่านั้น ผู้พัฒนาหวังเป็นอย่างยิ่งว่าโครงงานที่ได้พัฒนาขึ้นนี้จะช่วยในการพัฒนาสังคมและป้องกันเยาวชนให้ห่างไกลจากเว็บไซต์ที่ไม่เหมาะสมได้

5.3 ปัญหาอุปสรรคและแนวทางในการแก้ไข

1. เนื่องจากระบบใช้ทรัพยากรจำนวนมากเพื่อให้ได้ประสิทธิภาพสูงจำเป็น ต้องใช้อุปกรณ์ที่มีคุณภาพสูง

2. ระบบต้องการระบบเครือข่ายที่มีเสถียรภาพ ถ้าระบบเครือข่ายเกิดปัญหาจะทำให้โปรแกรมทำงานได้ไม่เต็มประสิทธิภาพ

3. การพิจารณาเว็บไซต์จากค่าน้ำหนักความไม่เหมาะสมเพียงอย่างเดียวอาจทำให้เกิดความผิดผลาดขึ้นได้ ควรทำการพิจารณาร่วมกับคุณสมบัติอื่นเช่น จำนวนเว็บเพจที่ค้นหาได้ของเว็บไซต์ ,จำนวนเว็บเพจที่ถูกพิจารณาแล้วว่าไม่เหมาะสมในเว็บไซต์ หรือ จำนวนเว็บเพจที่นำมาคำนวณหาค่าน้ำหนัก จะทำให้การจัดประเภทของเว็บไซต์มีความถูกต้องแม่นยำยิ่งขึ้น

5. ปัญหา spider trap หมายถึง การที่ URL หลาย URL อ้างอิงถึงหน้าเว็บเพจเดียวกันทำให้เกิดการวนลูปของการค้นหา ซึ่งสามารถแก้ไขได้ในระดับหนึ่งเท่านั้น

5.4 แนวทางการพัฒนาต่อ

1. ในอนาคตอาจมีการพัฒนาให้มีประสิทธิภาพมากยิ่งขึ้นและสามารถประยุกต์ใช้งานร่วมกับโครงการโปรแกรมคัดกรองข้อมูลเว็บ (Web content filtering) ซึ่งเป็นโปรแกรมที่ทำการคัดกรองเว็บไซต์อยู่บนเครื่องของผู้ใช้

2. พัฒนาเปลี่ยน Database Management System ให้มีประสิทธิภาพและสามารถรองรับการทำงานกับข้อมูลจำนวนมากได้เพื่อให้ระบบสามารถทำงานได้อย่างต่อเนื่องและมีประสิทธิภาพสูงสุด

3. การพัฒนาให้ระบบสามารถเรียนรู้รูปแบบโครงสร้างของเว็บไซต์ที่ไม่เหมาะสมซึ่งจะสามารถช่วยเพิ่มประสิทธิภาพการจำแนกให้สูงขึ้นได้

บรรณานุกรม

[1] H.M. Dietel, P.J. Dietel, J.P. Liperi, B.A. Wiedermann **Python How To Program**, Prentice Hall, Upper Saddle River, New Jersey, 2002.

[2] Rafael C. Gonzalez, Richard E. Woods **Digital Image Processing second edition**, Prentice Hall, Upper Saddle River, New Jersey, 2002.

[3] Guido van Rossum, Fred L. Drake **Python Tutorial Release 2.4**, Python Software Foundation

[4] จักรกฤษณ์ แสงแก้ว **การเขียนโปรแกรมภาษาไพธอนด้วยตนเอง,** สำนักพิมพ์ ส.ส.ท., กรุงเทพฯ, 2006

[5] Erwin Kreyszig **Advanced Engineering Mathematics Seventh Edition**, JOHN WILEY&SONS, INC., New York

[6] Gautam Pant, Padmini Srinivasan, Filippo Menczer "Crawling the Web"

[7] Liang K.M., Scott S.D., Waqas M. "Detecting Pornographic Image" , Asia Pacific Institute of Information Technology, Kuala Lumper, Malaysia .

[8] Feng Jiao, Wen Gao, Lijuan Duan, Guoqin Cui "Detecting Adult Image Using Multiple Features" , The Institute of Computing Technology, Beijing China, 2001.

[9] W.H Ho, P.A. Watters "Identifying and Blocking Pornographic Content"

[10] Yi Chan, Richard Harvey, Dan Smith "Building systems to block pornography", School of Information Systems, University of East Anglia, Norwich, UK, 1999.

[11] Qing-Fang Zheng, Wei Zeng, Gao Wen, Wei-Quing Wang "Shape-base Adult Images Detection" , The Institute of Computing Technology, Beijing China, 2004

[12] Christophe Garcia, Georgios Tziritas "Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis" , *IEEE Trans. Multimedia*, vol.1, no.3 ,PP. 264-277, September 1999.

[13] Lindsay I Smith "A tutorial on Principal Component Analysis" , February 26, 2002.

[14] Kardi Teknomo's Page "K-Mean Clustering Tutorial" [online]. Available: http://people.revoledu.com/kardi/tutorial/kmean/

[15] Kardi Teknomo's Page "K Nearest Neighbors Tutorial" [online]. Available: http://people.revoledu.com/kardi/tutorial/KNN/

ภาคผนวก

ภาคผนวก ก.

ตัวอย่างข้อมูลที่ถูกเก็บในฐานข้อมูล

ภาคผนวก ก.

ตัวอย่างข้อมูลที่ถูกเก็บในฐานข้อมูล

สำหรับภาคผนวกนี้เป็นตัวอย่างของข้อมูลที่เก็บลงในฐานข้อมูล ซึ่งประกอบด้วย 4 ตาราง สำคัญดังนี้

ตารางที่ ก. 1 ตัวอย่างข้อมูลในตาราง ALLURL

ID	URL	URL_ID	DATE_TIME	URL_TIMESTAMP	TEXT_WEIGHT
241	http://ztclips.com	0	12/1/2007 2:40	19/12/2006 23:57	12.5858
242	http://adultfriendfinder.com/go/g763989-ppc	3	12/1/2007 2:40	11/1/2007 19:39	14.898
243	http://ww2.bangmyhotwife.com/track/MTA0OTMwOjU6NTM	5	12/1/2007 2:40	11/1/2007 19:39	90.5955
244	http://www.free-gall.com	0	12/1/2007 2:40	11/1/2007 20:44	25.5184
245	http://ww2.brutalblowjobs.com/track/MTA0OTMwOjU6MTI	5	12/1/2007 2:40	11/1/2007 19:39	100
246	http://www.free-gall.com/ebony.html	3	12/1/2007 2:40	11/1/2007 20:44	55.6178
247	http://www.free-gall.com/penis/enlargement	3	12/1/2007 2:40	11/1/2007 20:44	32.746
248	http://ww2.wrongsideoftown.com/track/MTA0OTMwOjU6MzQ	5	12/1/2007 2:40	11/1/2007 19:39	59.8727
249	http://ww2.realbighooters.com/track/MTA0OTMwOjU6NTI	5	12/1/2007 2:40	11/1/2007 19:39	100
250	http://ww2.thebestpov.com/track/MTA0OTMwOjU6MzI	5	12/1/2007 2:40	11/1/2007 19:39	94.0696
251	http://ww2.gooeyholes.com/track/MTA0OTMwOjU6MjA	5	12/1/2007 2:40	11/1/2007 19:39	100
252	http://ww2.assplundering.com/track/MTA0OTMwOjU6OA	5	12/1/2007 2:40	11/1/2007 19:39	94.0696
253	http://ww2.fuckthebabysitter.com/track/MTA0OTMwOjU6NTY	5	12/1/2007 2:40	11/1/2007 19:39	29.1498
254	http://ww2.bjsandwich.com/track/MTA0OTMwOjU6MTA	5	12/1/2007 2:40	11/1/2007 19:39	83.8287
255	http://ww2.18inchesofpain.com/track/MTA0OTMwOjU6NTE	5	12/1/2007 2:40	11/1/2007 19:39	83.7682
256	http://ww2.fromasstomouth.com/track/MTA0OTMwOjU6MTY	5	12/1/2007 2:40	11/1/2007 19:39	100
257	http://ww2.plugherholes.com/track/MTA0OTMwOjU6MjQ	5	12/1/2007 2:40	11/1/2007 19:39	94.0696
258	http://ww2.mefuckyoulongtime.com/track/MTA0OTMwOjU6Mjl	5	12/1/2007 2:40	11/1/2007 19:39	100
259	http://ww2.gapeherass.com/track/MTA0OTMwOjU6MTg	5	12/1/2007 2:40	11/1/2007 19:39	100
260	http://ww2.semenslurpers.com/track/MTA0OTMwOjU6MjY	5	12/1/2007 2:40	11/1/2007 19:39	100
261	http://ww2.whiteboystomp.com/track/MTA0OTMwOjU6NTA	5	12/1/2007 2:40	11/1/2007 19:39	90.656
262	http://ww2.thebestlatinas.com/track/MTA0OTMwOjU6MzA	5	12/1/2007 2:40	11/1/2007 19:39	87.2423

ตารางที่ ก. 2 ตัวอย่างข้อมูลในตาราง IMAGE

ID	URL	DATE_TIME	IMAGE_WEIGHT
485	http://www.lesbiansex24-7.com/lesbian/sex_03.jpg	12/1/2007 3:02	50.7777
486	http://www.amateuralluregallery.com/wp-images/racheltn02.jpg	12/1/2007 3:05	89.1554
487	http://homemadesmutvideos.com/images/vid1g_10.gif	12/1/2007 2:56	Null
488	http://www.erotic-movies.ws/tour/picture/kim/kim009.jpg	12/1/2007 3:01	0
489	http://images.perfectgonzo.com/tt/tour/jacqueline_marcella/big.jpg	12/1/2007 3:02	0
490	http://holidayshaggers.com/bras4.jpg	12/1/2007 3:01	0
491	http://www.hardcorepartychicks.com/ads/partyhardcore.gif	12/1/2007 2:57	0
492	http://www.herfirst-kisses.com/banner/exmc.gif	12/1/2007 3:00	64.106
493	http://www.lesbiansex24-7.com/lesbian/sex_04.jpg	12/1/2007 3:03	98.6945
494	http://www.amateuralluregallery.com/wp-images/racheltn03.jpg	12/1/2007 3:05	94.713
495	http://www.big-tit-movies.net/images/doubledeckersandwich3/27.jpg	12/1/2007 3:03	83.1267
496	http://www.amateuralluregallery.com/wp-images/racheltn04.jpg	12/1/2007 3:05	95.8534
497	http://www.herfirst-kisses.com/i/topep.gif	12/1/2007 2:57	Null
498	http://www.amateuralluregallery.com/wp-images/alannatn01.jpg	12/1/2007 3:04	91.8889
499	http://www.hardcorepartychicks.com/images/13.jpg	12/1/2007 3:03	0
500	http://www.lesbiansex24-7.com/lesbian/sex_05.jpg	12/1/2007 3:03	77.7184
501	http://www.juicypornmovies.com/index_files/exit1_04.gif	12/1/2007 2:59	0
502	http://www.lesbiansex24-7.com/lesbian/sex_06.jpg	12/1/2007 3:04	0
503	http://homemadesmutvideos.com/images/vid1g_11.gif	12/1/2007 2:59	0
504	http://www.herfirst-kisses.com/i/bottitle.gif	12/1/2007 2:57	Null
505	http://images.perfectgonzo.com/tt/tour/jacqueline_marcella/1.jpg	12/1/2007 3:03	0
506	http://www.lesbiansex24-7.com/lesbian/sex_07.jpg	12/1/2007 3:04	83.6434
507	http://www.big-tit-movies.net/images/doubledeckersandwich3/31.jpg	12/1/2007 3:03	83.7624
508	http://404.webair.com/images/webair_top.gif	12/1/2007 2:58	Null
509	http://www.erotic-movies.ws/tour/picture/kim/kim027.jpg	12/1/2007 3:05	80.5389
510	http://www.hardcorepartychicks.com/images/14.jpg	12/1/2007 3:04	0
511	http://www.herfirst-kisses.com/i/dp-big.jpg	12/1/2007 3:02	0
512	http://404.webair.com/images/spacer.gif	12/1/2007 2:58	Null
513	http://www.juicypornmovies.com/index_files/009.gif	12/1/2007 3:05	0
514	http://www.erotic-movies.ws/tour/picture/kim/kim045.jpg	12/1/2007 3:06	75.6853
515	http://www.hardcorepartychicks.com/images/16.jpg	12/1/2007 3:03	0
516	http://images.perfectgonzo.com/tt/tour/jacqueline_marcella/2.jpg	12/1/2007 3:03	0
517	http://www.erotic-movies.ws/tour/picture/kim/kim020.jpg	12/1/2007 3:07	0
518	http://holidayshaggers.com/bod1.jpg	12/1/2007 3:03	0
519	http://thepornstararchive.adultvideoplanet.com/images/1.gif	12/1/2007 2:58	Null
520	http://www.big-tit-movies.net/images/doubledeckersandwich3/32.jpg	12/1/2007 3:04	0
521	http://homemadesmutvideos.com/images/vid1_12.jpg	12/1/2007 3:02	0
522	http://www.herfirst-kisses.com/i/raz.gif	12/1/2007 2:58	Null
523	http://images.perfectgonzo.com/tt/tour/jacqueline_marcella/3.jpg	12/1/2007 3:04	0
524	http://www.erotic-movies.ws/tour/picture/kim/kim037.jpg	12/1/2007 3:06	61.0875

ตารางที่ ก. 3 ตัวอย่างข้อมูลในตาราง PORN_**WEB**

ID	URL	IP_ADDRESS	DATE_TIME	TOTAL_WEIGHT	ISP_ID
200	http://www.thebestporn.com	66.115.191.31	31/1/2007 12:16	75.4192	22
201	http://www.seventhteen.com	66.28.176.121	31/1/2007 4:52	54.6816	1
202	http://www.foto-porno.ws	204.11.233.86	31/1/2007 12:16	33.3962	1
203	http://www.video-porno.nu	204.11.233.86	31/1/2007 12:17	21.8848	1
204	http://www.getyerrocksoff.com	198.65.152.37	31/1/2007 12:39	6.35841	1
205	http://www.kingssexsites.com	216.131.110.55	31/1/2007 4:59	16.9821	1
206	http://www.acmegirls.com	208.99.195.136	31/1/2007 12:42	21.4315	1
207	http://www.free-porn-index.com	216.17.106.208	31/1/2007 12:26	10.0689	1
208	http://www.adultpornguide.com	69.31.33.172	31/1/2007 12:31	34.2754	1
209	http://www.aboutmasturbation.com	207.210.231.42	31/1/2007 12:30	46.4327	1
210	http://www.saveonav.com	63.246.151.58	31/1/2007 12:21	11.0688	44
211	http://www.allysa.com	66.172.88.159	31/1/2007 12:26	7.97432	45
212	http://www.findmoreporn.com	64.255.22.53	31/1/2007 12:26	12.8934	46
213	http://www.free-sex-king.com	207.44.242.24	31/1/2007 12:26	34.0648	47
214	http://www.sexuall.org	65.99.249.230	31/1/2007 12:33	29.7055	1
215	http://www.youngsexclub.com	65.99.253.74	31/1/2007 12:23	8.13304	1
216	http://roughsexvids.com	64.72.121.253	31/1/2007 11:51	30.1573	1
217	http://www.xnostars.com	207.210.90.189	31/1/2007 12:36	35.7455	1
218	http://www.matureplace.com	66.172.88.132	31/1/2007 12:24	16.6197	45
219	http://www.relatos-putas.com	72.232.137.178	31/1/2007 12:16	3.58664	1
220	http://mikgalleries.com	207.226.180.18	31/1/2007 11:55	39.565	1
221	http://gaietymaster.com	216.195.51.135	31/1/2007 4:53	0.869307	1
222	http://www.jays-xxx-links.com	66.115.140.219	31/1/2007 12:36	42.1978	22
223	http://www.teenlesbiansporn.com	66.28.176.121	31/1/2007 12:26	53.4749	1
224	http://www.celebrityvideozone.com	209.200.55.177	31/1/2007 12:27	23.7965	1
225	http://www.xxxthumbs4free.com	66.250.223.102	31/1/2007 12:39	47.734	1
226	http://www.purecelebsite.com	82.208.62.110	31/1/2007 12:44	26.5694	21
227	http://www.sweetestteenie.com	194.126.193.175	31/1/2007 12:27	32.6323	17
228	http://www.amateuralluregallery.com	64.237.47.24	31/1/2007 11:59	62.5666	1
229	http://www.celebritymovieblog.com	82.208.60.201	31/1/2007 12:39	38.0529	21
230	http://www.greenthumbstgp.com	69.31.38.70	31/1/2007 12:27	26.256	1
231	http://www.bondagepoint.com	82.208.63.107	31/1/2007 11:51	21.6467	21
232	http://www.best-deepthroat-blowjobs.co.uk	66.55.73.6	31/1/2007 12:31	20.0489	41
233	http://www.freegally.com	66.230.177.130	31/1/2007 12:46	67.8483	1
234	http://www.teenporncasting.com	209.200.18.41	31/1/2007 12:30	2.69852	1
235	http://www.sexy-photos.net	81.0.235.170	31/1/2007 12:39	52.3276	48
236	http://www.celebrityacademy.com	69.25.180.228	31/1/2007 12:42	19.9702	1
237	http://www.tgpmadness.com	66.28.176.121	31/1/2007 12:42	22.0918	1
238	http://www.milf-seeker-search.com	66.230.138.29	31/1/2007 12:41	60.961	5
239	http://www.plumpster.net	65.98.60.114	31/1/2007 12:06	9.55729	1
240	http://deepthroatguide.com	195.56.55.72	31/1/2007 12:49	39.204	37
241	http://www.kidsnet.com	64.182.115.168	31/1/2007 12:31	0.328316	1
242	http://www.sexwp.com	72.11.136.122	31/1/2007 12:33	19.4722	1

ตารางที่ ก. 4 ตัวอย่างข้อมูลในตาราง ISP

ID	NAME	COUNTRY	PHONE	E_MAIL	IP_RANGE
223	FLATBOX-MNT	NL	+31 70 3589165	ripe@flatbox.nl	62.204.64.0 - 62.204.67.255
224	EWEKA-MNT	NL	+31 72 5640406	ripe@eweka.nl	81.171.32.0 - 81.171.127.255
225	ABRARED-MNT	es	+34 639 652 586	norberto_navarro@hotmail.com	80.81.124.64 - 80.81.124.127
226	AS3340-MNT	HU	+36 1 8144351	----	195.56.55.48 - 195.56.55.63
227	ABOVENET-P	DE	+49 6341 9284 0	dausch@megaspace.de	82.98.201.0 - 82.98.201.255
228	Content Broadcast	US	----	arin@contentbroadcast.com	64.185.224.0 - 64.185.239.255
229	REDES-MNT	ES	+34 912127620	sebastian.muriel@red.es	194.69.254.0 - 194.69.254.255
230	neoworld-mnt	NL	+48 71 344 8838	----	194.126.172.0 - 194.126.175.255
231	Microsoft Corp	US	----	iprrms@microsoft.com	207.68.128.0 - 207.68.207.255
232	EWEKA-MNT	NL	+31 773 969 975	----	195.74.65.0 - 195.74.65.255
233	LNC-MNT	DE	+49 1805 47328638	----	83.133.96.0 - 83.133.127.255
234	America Online	US	----	domains@aol.net	152.163.0.0 - 152.163.255.255
235	AS5486-MNT	IL	+972 39 399 703	wan@zahav.net.il	213.8.193.0 - 213.8.193.255
236	REDIRIS-NMC	ES	+34 913495058	----	193.147.89.0 - 193.147.89.255
237	Level 3 Communications, Inc.	US	----	arin-contact@genuity.com	8.0.0.0 - 8.255.255.255
238	TRUESERVER-MNT	NL	+31 20 30 59 750	noc@trueserver.nl	85.255.208.0 - 85.255.223.255
239	OCOM-MNT	NL	+31 20 3162880	----	82.192.69.0 - 82.192.69.127
240	OCOM-MNT	NL	+31 20 3162880	----	83.149.98.0 - 83.149.98.255
241	OCOM-MNT	NL	+31 20 3162880	----	85.17.42.0 - 85.17.42.255
242	XS4ALL-MNT	NL	+31 20 3987654	----	82.94.237.224 - 82.94.237.255
243	WZNET-MNT	NL	+31 612 253 464	noc@webazilla.com	88.85.74.128 - 88.85.74.255
244	Hostway Corporation	US	----	noc@hostway.com	66.232.128.0 - 66.232.159.255
245	MNT-REASONNET	PH	+63 754327753	----	85.92.152.0 - 85.92.159.255
246	IPPARTNER-MNT	DE	+49 911 30950 000	----	80.190.203.0 - 80.190.203.63

ภาคผนวก ข.

ตัวอย่างข้อมูลที่ได้จากการทดลอง

ภาคผนวก ข.

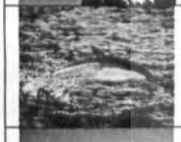
ตัวอย่างข้อมูลที่ใช้ในการทดลอง

สำหรับภาคผนวกนี้เป็นตัวอย่างของข้อมูลใช้ในการทดลอง ซึ่งประกอบด้วยข้อมูลของเนื้อหาภายในเว็บไซต์ทั้ง 10 คุณสมบัติ และข้อมูลของภาพทั้ง 6 คุณสมบัติ

ตารางที่ ข.1 คุณสมบัติของเนื้อหาภายในเว็บเพจของข้อมูลตัวอย่างทั้ง 10 คุณสมบัติ

URL	คุณสมบัติ									
	Img Tag	Meta Tag	คำหยาบคาย	Script Tag	Link	Img Link	Text Link	Non-Relative Link	Relative Link	Param Tag
http://www.spunklords.com	119	0	2	1	224	113	111	80	144	0
http://www.free-pornstar-pictures.biz	4	2	13	0	17	4	13	12	5	0
http://www.plainpost.com	62	5	8	1	225	1	224	27	198	0
http://www.marissas.com	1	6	5	0	350	1	349	191	159	0
http://www.2freexxx.com	0	5	26	0	93	0	93	80	13	0
http://candyclips.net	21	1	1	0	38	19	19	38	0	0
http://debauchery.com	8	2	20	0	246	7	239	245	1	0
http://www.throatsex.com	37	1	1	0	14	9	5	14	0	0
http://www.pornopat.com	27	7	7	1	82	26	56	77	5	0
http://www.erotiqlinks.com	7	3	11	4	543	6	537	324	219	0
http://www.seventhteen.com	15	5	21	2	379	14	365	88	291	0
http://www.freepornoz.com	4	2	13	3	181	4	177	181	0	0
ttp://www.xnostars.com	42	5	14	3	287	17	270	255	32	0
http://www.mmm100.com	117	5	36	9	281	115	166	29	252	0
http://www.netsexplaces.com	4	7	5	3	262	1	261	248	14	0
http://www.sultans-porn.com	78	12	16	9	305	5	300	197	108	0
http://www.websex4u.net	216	18	15	5	312	105	207	98	214	0
http://www.milfshake.com	3	8	29	0	54	2	52	54	0	0
http://www.joggs.com	36	10	10	3	277	22	255	244	33	0
http://www.quality-adult-sex.com	24	11	32	2	181	8	173	160	21	0
http://www.puritanas.com	88	15	19	9	308	60	248	252	56	0
http://www.huntcelebs.com	18	2	4	0	900	9	891	26	874	0
http://www.jrmovies.com	55	2	4	0	84	28	56	3	81	0
http://www.xnxx.com	144	4	9	3	524	16	508	293	231	0
http://www.tits-bigtits.com	131	5	19	1	324	116	208	176	148	0
http://www.sexwp.com	1	9	7	0	14	1	13	14	0	0
http://www.riaps.com	6	1	2	2	433	6	427	76	357	0
http://www.usshemales.com	86	6	15	4	274	22	252	60	214	0
http://www.dildo-x.com	26	6	25	1	13	1	12	4	9	0
http://www.teenjizz.net	71	2	0	1	87	63	24	10	77	0
http://www.bestporn4u.com	175	9	14	8	222	10	212	138	84	0
http://www.sex--sex.us	2	8	2	0	198	2	196	193	5	0
http://www.foto-porno.ws	0	3	8	4	8	0	8	2	6	0
http://www.1gayporn.com	170	4	5	56	342	10	332	252	90	0

ตารางที่ ข.2 คุณสมบัติของภาพของข้อมูลตัวอย่างทั้ง 6 คุณสมบัติ

ภาพ	คุณสมบัติ					
	พื้นที่สีผิว (%)	พื้นที่สีผิวขนาดใหญ่ที่สุด (%)	จำนวน region ของสีผิว (ชิ้น)	ความสูงของพื้นที่ผิว (%)	ความกว้างของพื้นที่ผิว (%)	พื้นที่ผิวขนาดใหญ่เทียบกับพื้นที่ผิวทั้งหมด(%)
	16.5333	3.4667	7	38.6667	37.0	20.9678
	9.6	7.6	3	42.6667	45.0	79.1667
	3.6667	2.0152	2	24.2424	14.0	54.9587
	8.7	6.0375	4	48.75	40.0	69.3965
	1.3067	1.3067	1	8.0	28.0	100.0
	4.394	1.9393	3	34.8485	11.0	44.1379
	11.1867	11.1867	1	38.6667	99.0	100.0
	11.96	8.87	4	38.0	78.0	74.16
	9.6933	5.1333	2	73.3333	22.0	52.9574
	21.5875	21.5875	1	87.5	52.0	100.0
	21.08	10.5866	4	56.0	61.0	50.2214
	21.92	7.0267	9	56.0	28.0	32.0559