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

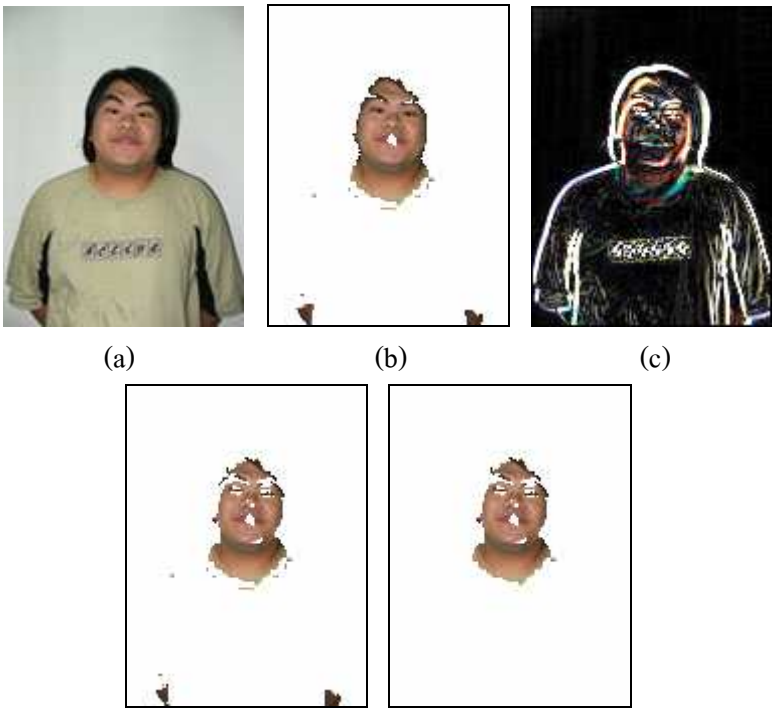
โครงงานนี้ได้ใช้ภาษาไพธอนในการพัฒนาส่วนของโปรแกรมตรวจสอบภาพอนาจรตามขั้นตอนที่ได้กล่าวไว้ และทดสอบการทำงานของโปรแกรมที่ได้พัฒนาขึ้น ดังนี้

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)



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
- ทำการพิจารณาหาพื้นที่ผิวที่มีความต่อเนื่องกันซึ่งมีขนาดใหญ่ที่สุดในภาพ และคำนวณเปอร์เซ็นต์ของพื้นที่ที่ใหญ่ที่สุดต่อจำนวนพิกเซลทั้งหมดของภาพ
- 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 เรื่องกระบวนการวิเคระห์ส่วนประกอบ 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

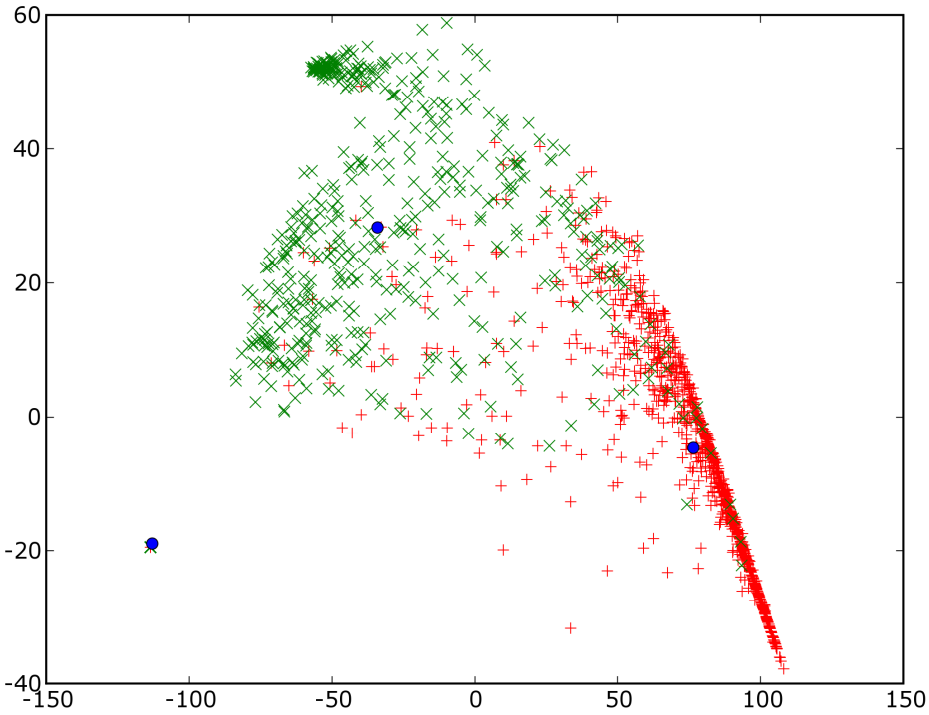
- เลือกคุณสมบัติที่มีความสำคัญสูงที่สุด 2 คุณสมบัติที่ทำให้สูญเสียข้อมูลน้อยที่สุด จากค่าไอแกนแวลูส์ และไอแกนเวกเตอร์ที่คำนวณได้ ทำการเลือกไอแกนเวกเตอร์ที่มีค่าไอแกนแวลูส์สูงที่สุด 2 อันดับแรก คือ 6.81879802e+03 และ 5.16192545e+02 ซึ่งมีไอแกนเวกเตอร์คือ [0.37311171 , 0.37007975 , 0.03439508 , 0.50656072

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>

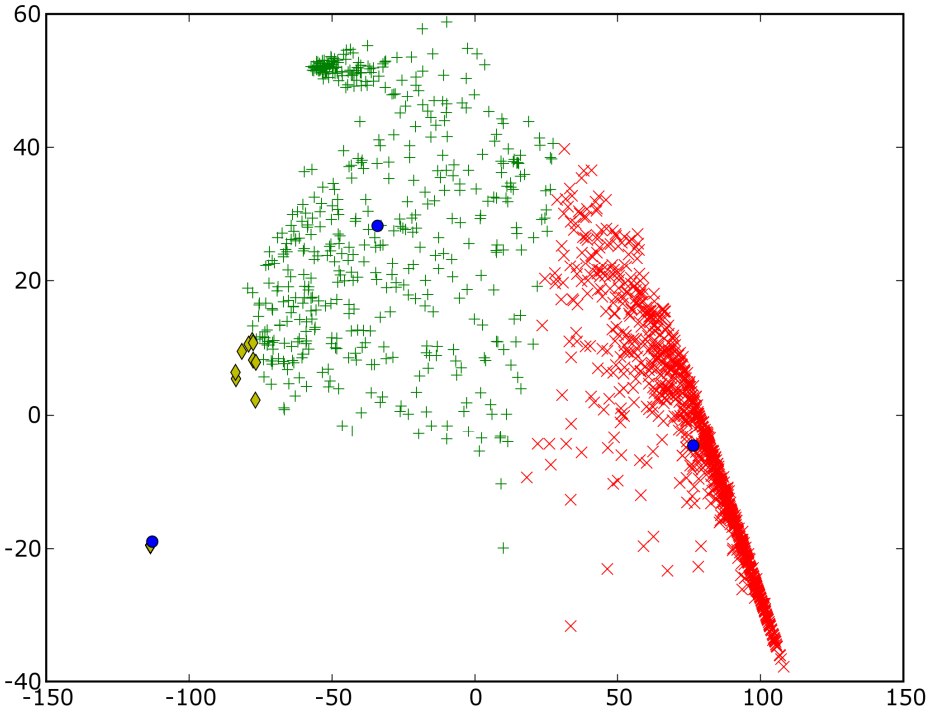
จากการทดลองพัฒนาโปรแกรมการจัดประเภทของภาพด้วยคุณสมบัติด้วยภาษาไพธอนจะได้ผลดังรูปที่ 4.2



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



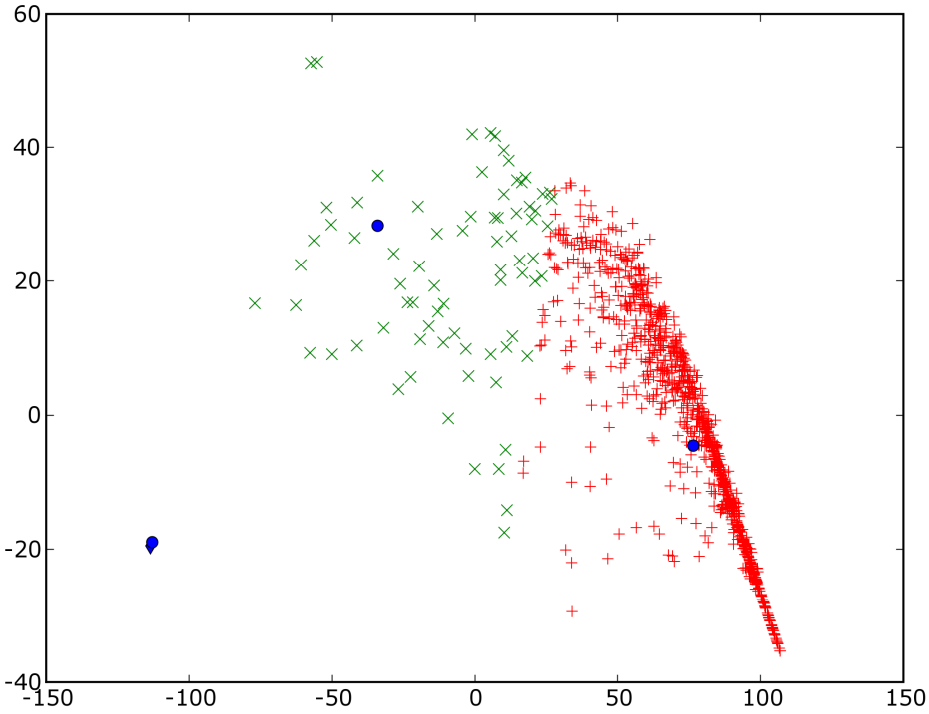
รูปที่ 4.3 ผลการทดลองการจัดกลุ่มข้อมูลของตัวอย่างภาพ

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

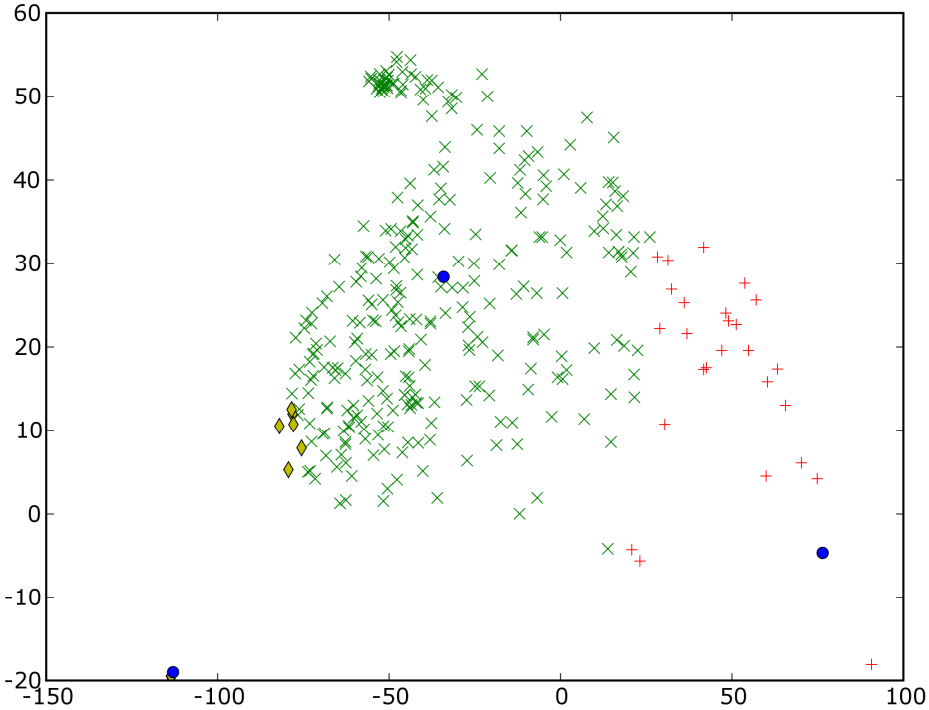
ตารางที่ 4.2 ประสิทธิภาพของการจัดกลุ่มข้อมูลของตัวอย่างภาพ

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

<u>การทดลองที่ 2</u>



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

<u>ผลการทดสอบประสิทธิภาพ</u>

โปรแกรมส่วนของ Image Detection สามารถตรวจสอบและคำนวณค่าน้ำหนักของรูปได้โดยใช้เวลา เฉลี่ย 1.724 วินาทีต่อภาพ

4.3 การทดสอบโปรแกรมการวิเคราะห์เนื้อหาในเว็บไซต์

การตรวจสอบเนื้อหาของเว็บเพจจะประกอบด้วยขั้นตอนการตรวจสอบคุณสมบัติของเว็บเพจและการจัดประเภทของเนื้อหาเว็บเพจด้วยคุณสมบัติ ซึ่งใช้หลักการ Regular Expression เข้ามาช่วยในการตรวจสอบ ขั้นตอนการทดลองการทำงานของโปรแกรมที่ได้พัฒนาขึ้น มีดังนี้

4.3.1 การตรวจสอบคุณสมบัติของเนื้อหาในเว็บเพจ

<u>วิธีการทดลอง</u>

- 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

<u>ผลการทดลอง</u>

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

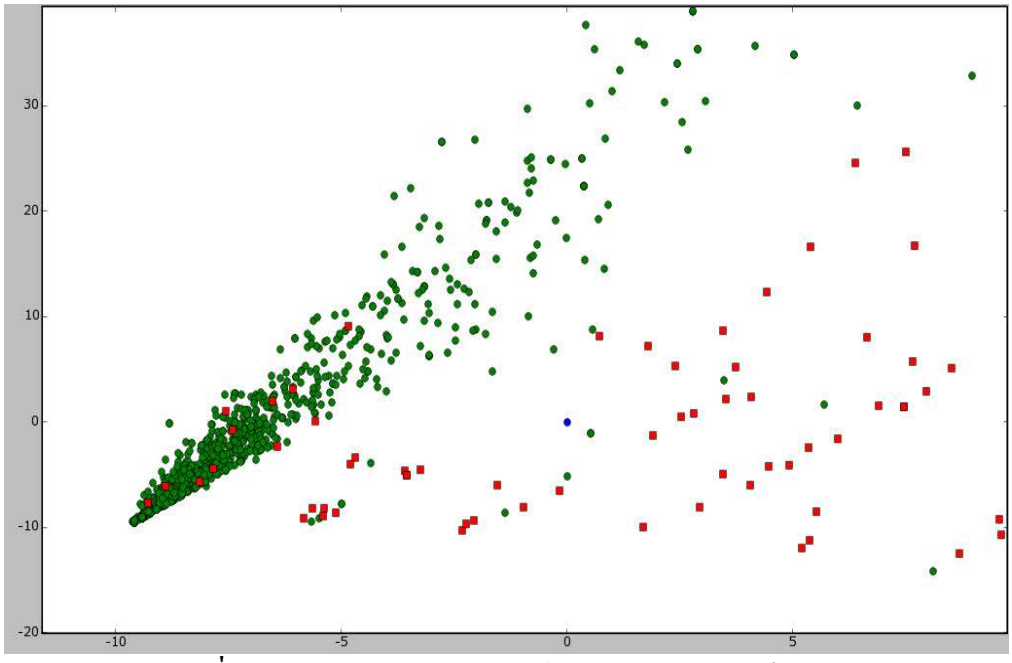
จากการทดลองการตรวจสอบคุณสมบัติของเนื้อหาในเว็บเพจจะเห็นว่าบางคุณสมบัติไม่สามารถใช้เป็นตัวระบุได้ว่าเว็บเพจนั้นเป็นเว็บอนาจารได้อย่างชัดเจน ดังนั้นจึงต้องมีการนำคุณสมบัติของเว็บเพจทั้ง 10 ประการมาผ่านหลักการการวิเคราะห์ส่วนประกอบ (Principle Component Analysis) โดยนำคุณสมบัติทั้ง 10 มาพิจารณาร่วมกันเพื่อให้สามารถระบุว่าเว็บนั้นเป็นเว็บอนาจารได้อย่างถูกต้อง

<u>วิธีการทดลอง</u>

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

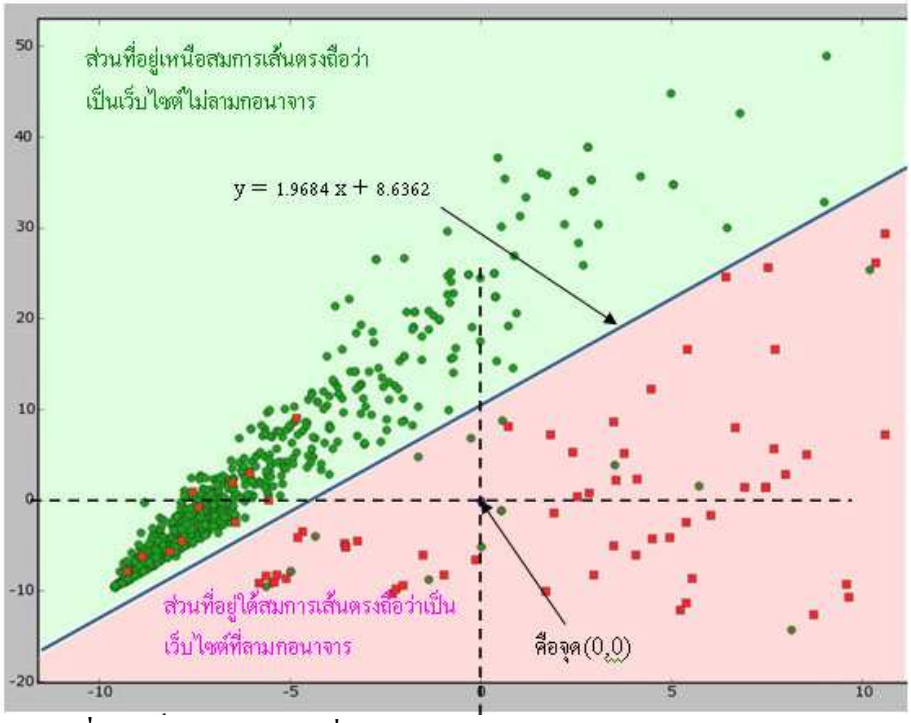
<u>ผลการทดลอง</u>

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

<u>ผลการทดสอบประสิทธิภาพ</u>

โปรแกรมส่วนของการตรวจสอบเนื้อหาของเว็บเพจสามารถตรวจสอบคุณสมบัติและคำนวณค่าน้ำหนักของเว็บเพจได้โดยใช้เวลาเฉลี่ย 1.560 วินาทีต่อเว็บเพจ

4.4 การทดสอบประสิทธิภาพระบบ

<u>ผลการทดสอบประสิทธิภาพการใช้ทรัพยากรของระบบ</u>

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

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

<u>ผลการทดสอบประสิทธิภาพโดยรวม</u>

จากกลุ่มตัวอย่างสามารถคำนวณประสิทธิภาพและความผิดพลาดได้ดังนี้

1. โปรแกรมส่วนของการตรวจสอบรูปภาพมีอัตราความถูกต้อง 92.7% และอัตราความผิดพลาด 7.3%
2. โปรแกรมส่วนของการตรวจสอบเนื้อหาของเว็บเพจมีอัตราความถูกต้อง 97.25% และอัตราความผิดพลาด 2.75%
3. โปรแกรมทั้งหมดมีอัตราความถูกต้อง 97.6% และอัตราความผิดพลาด 2.4%