# บทที่ 1
# บทนำ

## 1.1 ที่มาและความสำคัญของปัญหา

ปัจจุบันการโฆษณาบนเว็บไซต์ (Online advertising) มีแนวโน้มเพิ่มมากขึ้น เนื่องจากการพัฒนาของผู้ใช้งานอินเทอร์เน็ตเพิ่มขึ้น ดังนั้นการโฆษณาจึงต้องมีการปรับเปลี่ยนตามเพื่อเพิ่มการตลาดให้ทันสมัยและรองรับการใช้งานตามความต้องการของผู้ใช้ ซึ่งประเภทของโฆษณาบนเว็บไซต์จะเป็นข้อความโฆษณาที่เห็นทั่วไปตามเว็บไซต์ เช่น โปรแกรมที่ช่วยในการสืบค้นหาข้อมูล (Search engine) การโฆษณาตามระบบเครือข่าย ระบบการจัดส่งจดหมายข่าวสารทางอิเล็กทรอนิกส์ที่เป็นจดหมายโฆษณา (Email Marketing) นอกจากนี้ยังมีการโฆษณาในรูปแบบวิดีโอ (Video online) ซึ่งจะมีรูปแบบเหมือนการโฆษณาตามโทรทัศน์ ถ้าสนใจสามารถคลิกเข้าสู่เว็บไซต์นั้นได้โดยตรง นอกจากนี้การโฆษณายังมีอยู่ส่วนของจดหมายอิเลกทรอนิกส์และเบอร์โทรศัพท์ ซึ่งจะสามารถบ่งบอกถึงการทำธุรกิจได้มากขึ้น

สำหรับแนวคิดในการบล็อก (Block) รูปภาพโฆษณานั้นมีผู้เริ่มทำการศึกษาและวิจัยตั้งแต่ปี ค.ศ.1999 แต่เนื่องจากเว็บไซต์มีการปรับเปลี่ยนไปตามยุคสมัยและรูปแบบการโฆษณามีความหลากหลายเพิ่มมากขึ้น งานวิจัยที่เกี่ยวข้องส่วนใหญ่จะใช้วิธีการหาคุณลักษณะที่แตกต่างของภาพที่เป็นโฆษณากับภาพที่ไม่เป็นโฆษณาแล้วนำมาวิเคราะห์หาคุณลักษณะเด่น (Feature) [1] ซึ่งคุณลักษณะเด่นที่ใช้ในอดีตอาจไม่ครอบคลุมสำหรับเว็บไซต์ปัจจุบันเพราะลักษณะการทำเว็บไซต์มีการพัฒนาอย่างรวดเร็ว ดังนั้นการบล็อกรูปภาพที่เป็นโฆษณาจึงมีความยุ่งยากและซับซ้อนมากขึ้นตามไปด้วย

ปัญหาของภาพโฆษณาบนเว็บไซต์และประโยชน์ของงานวิจัย

1.1.1 ลดการใช้แบนด์วิท (Bandwidth) ในการส่งผ่านข้อมูลต่างๆ บนระบบออนไลน์ เพราะภาพโฆษณาส่วนใหญ่ปรากฏอยู่ทั่วไปในหน้าเว็บไซต์และเป็นภาพกราฟิก ทำให้การทำงานในส่วนอื่นช้าลง ดังนั้นงานวิจัยนี้จะบล็อกภาพโฆษณาเหล่านั้นเพื่อทำให้การทำงานเร็วขึ้น เป็นการลดการใช้แบนด์วิทและเพิ่มความเร็วในการใช้งาน

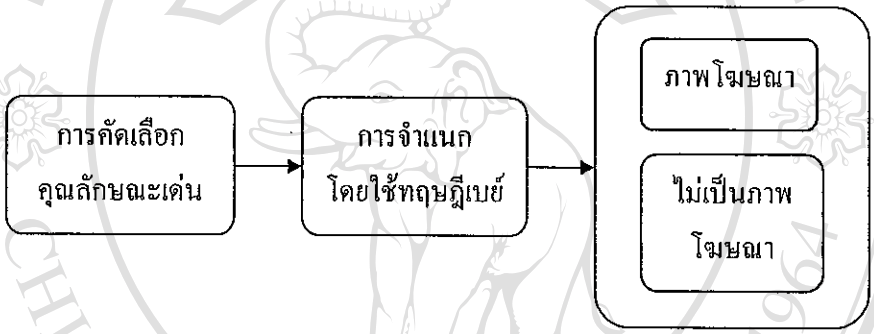
1.1.2 โฆษณาบางอย่างมีแฟ้มข้อมูลที่แอบแฝงมากกว่าการโฆษณา เช่น ถ้าผู้ใช้คลิกเข้าไปในรูปที่เป็นโฆษณาหรือมีการดาวน์โหลด (Download) ข้อมูลซึ่งอาจมีแฟ้มข้อมูลที่ไม่เป็นประโยชน์แฝงตัวมาด้วย ดังนั้นถ้าโฆษณาเป็นลักษณะนี้งานวิจัยนี้จะมีประโยชน์ที่สามารถกรอง

ข้อมูลที่เข้ามาให้มีความปลอดภัยมากยิ่งขึ้น

1.1.3 รูปแบบการโฆษณาและการเขียนโปรแกรมบนเว็บไซต์มีความหลากหลายมากขึ้น เพราะฉะนั้นงานวิจัยนี้จะต้องปรับปรุงอัลกอริทึม (Algorithm) เพื่อให้เข้ากับสถานการณ์ปัจจุบัน และให้สามารถทำงานได้อย่างมีประสิทธิภาพ

**1.2 แนวทางการแก้ปัญหา**

งานวิจัยนี้ได้นำเสนอวิธีการคัดเลือกคุณลักษณะเด่นที่สามารถบล็อกภาพโฆษณาประเภทแบนเนอร์ (Banner)ได้โดยใช้อัลกอริทึมการจำแนกประเภทแบบเบย์ ซึ่งสามารถจำแนกรูปภาพที่เป็นโฆษณาและไม่เป็นโฆษณาบนเว็บเพจได้




**รูปที่ 1.1** ระบบรวมที่แสดงขั้นตอนหลักการทำงานของงานวิจัย

**1.3 วัตถุประสงค์ของการวิจัย**

1.3.1 ศึกษาการทำงานของภาพโฆษณาบนเว็บไซต์ประเภทแบนเนอร์และนำเสนอคุณลักษณะเด่นใหม่ที่สามารถแยกความแตกต่างของภาพที่ต้องการและภาพที่ไม่เป็นโฆษณา แล้วทำการบล็อกโดยที่ไม่ให้เว็บเพจแสดงภาพที่เป็นโฆษณาประเภทแบนเนอร์ออกมาเมื่อมีการเข้าสู่เว็บไซต์นั้น

1.3.2 ศึกษาและนำเสนอแนวคิดรวมถึงวิธีการสำหรับการนำคุณลักษณะเด่นที่คิดขึ้นใหม่มาใช้ร่วมกับคุณลักษณะเด่นเดิมเพื่อเพิ่มประสิทธิภาพของงานวิจัย

1.3.3 ศึกษาการทำงานโดยใช้การจำแนกสำหรับทฤษฎีการเรียนรู้ (Learning) ที่นำมาใช้กับงานวิจัย โดยอัลกอริทึมนี้จะทำการฝึกฝนกลุ่มข้อมูล (Train) และทำการทดสอบข้อมูล (Test) เพื่อให้สามารถทำการจำแนกรูปภาพและทำการบล็อกภาพที่เป็นโฆษณาประเภทแบนเนอร์ได้อย่างถูกต้อง

1.2.4 ศึกษาระบบการทำงานและการส่งผ่านข้อมูลระหว่างเซิฟเวอร์ (Server) ดีเอ็นเอสเซิฟเวอร์ (DNS Server) พร็อกซี่เซิฟเวอร์ (Proxy Server) และระบบเครือข่าย เนื่องจากงานวิจัย

นี้จะทำการบล็อกภาพที่เป็นโฆษณาบริเวณส่วนพร็อกซี่เซิฟเวอร์ก่อนที่จะส่งผ่านข้อมูลไปยังเครื่องลูกข่ายและเข้าสู่ระบบเครือข่าย

**1.4 ขอบเขตของการศึกษาวิจัย**

1.4.1 การบล็อกภาพที่เป็นโฆษณาลักษณะแบนเนอร์แบ่งเป็น 2 ประเภทดังนี้

ประเภทที่ 1 แบบสแตติกเว็บเพจ (Static web page) เป็นภาพโฆษณาที่มีขนาดและตำแหน่งที่กำหนดชัดเจนและนำมาวางไว้ในตำแหน่งต่างๆ ของหน้าเว็บไซต์ ส่วนใหญ่มีลักษณะเป็นภาพสัญลักษณ์ของสินค้าซึ่งเมื่อมองเห็นภาพจะทราบได้ทันทีว่าเป็นภาพโฆษณาสินค้านั้น

ประเภทที่ 2 แบบไดนามิกเว็บเพจ (Dynamic web page) เป็นภาพโฆษณาที่มีขนาดและตำแหน่งเช่นกัน โดยทุกๆภาพมีการเรียกแฟ้มข้อมูลจากภาพที่เก็บไว้ในฐานข้อมูลหรือภาพลักษณะสแตติกเพื่อนำมาแสดงในตำแหน่งที่กำหนดโดยให้มีลักษณะสุ่มหรือเปลี่ยนภาพที่แสดงไปเรื่อยๆ แต่ภาพจะอยู่ในตำแหน่งเดิม

1.4.2 สร้างโปรแกรมคอมพิวเตอร์ที่เป็นอัลกอริทึม (Algorithm) ในลักษณะเป็นตัวกรองข้อมูลไว้ที่พร็อกซี่เซิฟเวอร์สำหรับการบล็อกภาพ

1.4.3 สร้างเครื่องแม่ข่าย เครื่องลูกข่ายและจำลองระบบเน็ตเวิร์คที่จะใช้ทำการทดสอบงานวิจัยที่ได้ทำการศึกษา

1.4.4 ทำการทดลองและทดสอบผลกับระบบการทำงานสำหรับใช้ในงานวิจัย เพื่อทำการประเมินประสิทธิภาพของงานวิจัยโดยการทดลองบล็อกภาพที่พร็อกซี่เซิฟเวอร์ผ่านอัลกอริทึมที่ทำขึ้นแล้วเปรียบเทียบผลการทดลองโดยวัดประสิทธิภาพของงานวิจัยกับผลระบบการทำงานจริง

1.4.5 สรุปผลการทดลองของงานวิจัยนี้และวิเคราะห์ผลที่ได้จากวิธีการที่ได้นำเสนอ

**1.5 ประโยชน์ที่คาดว่าจะได้รับ**

1.5.1 สามารถแยกภาพที่ไม่เป็นโฆษณากับภาพที่เป็นโฆษณาได้โดยอ่านข้อมูลจากแฟ้มข้อมูลภาษาเอชทีเอ็มแอล

1.5.2 สามารถลดการใช้แบนด์วิทที่ใช้ในระบบเน็ตเวิร์ค ทำให้เพิ่มความเร็วในการโหลดข้อมูลมากขึ้น

1.5.3 สามารถหาอัลกอริทึมที่เหมาะสมสำหรับการบล็อกภาพโฆษณาประเภทแบนเนอร์เพื่อทำให้งานวิจัยเกิดประโยชน์สูงสุด

1.5.4 สามารถคิดคุณลักษณะเด่นขึ้นมาใหม่เพื่อเพิ่มประสิทธิภาพในการบล็อกให้มากขึ้น

1.5.5 สามารถนำโปรแกรมคอมพิวเตอร์ที่ได้ไปใช้พัฒนาและปรับปรุงต่อไปในอนาคต

**1.6 สรุปสาระสำคัญจากเอกสารที่เกี่ยวข้อง**

สำหรับงานวิจัยที่ผ่านมาได้มีการใช้คุณลักษณะเด่นหลายประเภทที่สามารถจำแนกภาพได้ เช่น ภาพที่มีการเคลื่อนไหว (Animation) ภาพแฟลช (Flash) คือโปรแกรมสำหรับสร้างข้อความ ภาพ เสียงที่เคลื่อนไหวและโต้ตอบกับผู้ชมได้ เป็นโปรแกรมที่ได้รับความนิยมสูงมาก มีการนำมาใช้ในเว็บไซต์อย่างแพร่หลาย เช่น ข้อความ ภาพเคลื่อนไหว เสียง เกมส์ ภาพยนตร์สั้นๆ การ์ตูน ภาพลักษณะนี้ส่วนใหญ่มักจะนำมาเป็นภาพที่เป็นโฆษณา เพราะต้องการให้สินค้าหรือลักษณะการโฆษณามีจุดเด่น น่าดึงดูดใจที่จะเข้าไปชมเว็บไซต์ รวมถึงการเพิ่มความเร็วในการโหลดข้อมูลและการแสดงผล (Render) เว็บไซต์ ส่วนใหญ่มีชนิดแฟ้มข้อมูลประเภท.GIF , ประเภท .SWF และประเภท.SVG เป็นต้น สำหรับการใช้คุณลักษณะเด่นนี้จะทำการตรวจสอบในแฟ้มข้อมูลภาษาเอชทีเอ็มแอลเช่นกัน โดยตรวจสอบที่แท็กของรูปภาพ คำสั่งที่ใช้คือแท็ก <IMG>

งานวิจัยในอดีตจะนิยมใช้คุณลักษณะเด่นสำหรับใช้ในการจำแนกภาพที่เป็นโฆษณาดังนี้

1.6.1 ยูอาร์แอล (Uniform Resource Locator) เป็นการบ่งบอกตำแหน่งของข้อมูลในเวิลด์ไวด์เว็บ (www) จะพบในเว็บบราวเซอร์เมื่อเข้าสู่เว็บไซต์

1.6.2 ค่าอัตราส่วนของภาพ (Ratio) เป็นอัตราส่วนของรูปภาพที่เป็นโฆษณา มีขนาดเป็นพิกเซล (Pixel)

1.6.3 คำ (Word) ที่อยู่ในแท็ก <IMG SRC= ../คำ..> เช่น ads , ad banner, ad frame, advertise , ad click เป็นต้น

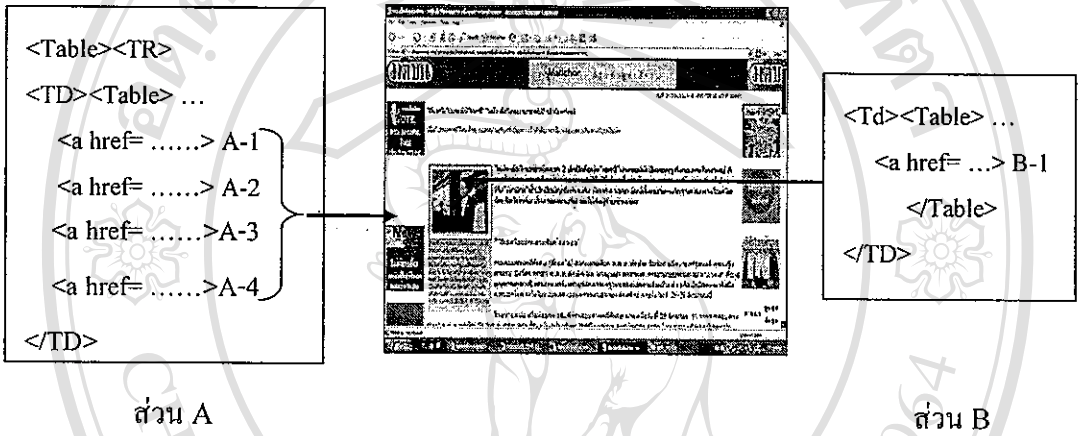
1.6.4 คำที่อยู่ในแท็ก <alt> เมื่อใช้เมาส์ไปวางตรงตำแหน่งที่เป็นภาพนั้นจะมีคำที่น่าจะเป็นโฆษณาขึ้นมา เช่น Contact us , Advertisement เป็นต้น

แท็ก <IMG> มีความหมายคือบอกบราวเซอร์ให้รู้ว่าข้อมูลต่อไปนี้เป็นชื่อแฟ้มข้อมูลรูปภาพที่จะแสดงในเว็บไซต์ และ “SRC=” ตามด้วยชื่อแฟ้มข้อมูลรูปภาพที่ต้องการแสดงจะระบุที่เก็บแฟ้มข้อมูลที่มีชนิดเป็นแฟ้มข้อมูลรูปเช่น .gif , .jpg , .swf ,.svg เป็นต้น ดังนั้นการบล็อกโฆษณาที่ผ่านมาจึงใช้คุณลักษณะเด่นนี้

ข้อความ(Word)ที่ปรากฏอยู่เมื่อใช้เมาส์เลื่อนไปยังตำแหน่งรูปภาพเป็นคุณลักษณะเด่นอีกประเภทหนึ่งที่มีผู้ใช้สำหรับแยกภาพ ข้อความลักษณะนี้จะอยู่ติดหรืออยู่บริเวณรูปภาพ ซึ่งในส่วนของโปรแกรมเป็นคำสั่งที่กำหนดให้มีการใช้ข้อความซ้ำๆกันที่หมายถึงการโฆษณา เช่น Click , Our sponsor, Contact us, Click here เป็นต้น และในการบล็อกจะทำการตรวจสอบว่ามีคำหรือข้อความเหล่านี้อยู่ในตำแหน่งหรือบริเวณใกล้ๆ กับภาพที่ทำการวิเคราะห์อยู่หรือไม่ ถ้ามีการใช้คำเหล่านี้ก็มีความเป็นไปได้ที่ภาพเหล่านั้นจะเป็นโฆษณา

นอกจากนี้ยังมีคุณลักษณะเด่นที่ใช้ได้ผลดีคือโครงสร้างตาราง (HTML Table Tree) [6]

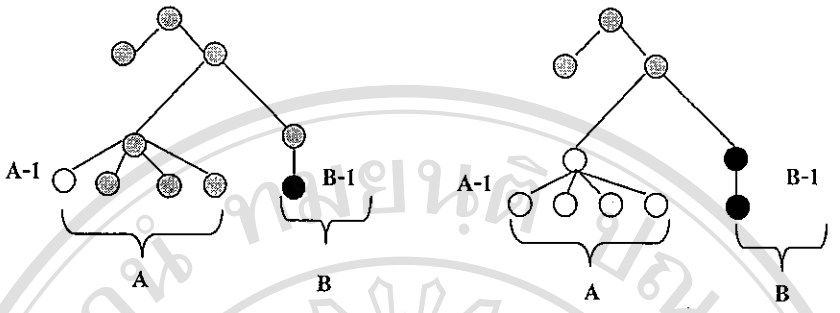
เนื่องจากในการเขียนเว็บเพจ การสร้างตารางเป็นส่วนประกอบหลักสำหรับการวิเคราะห์โครงสร้างภายในของแต่ละส่วนโดยการวิเคราะห์จากแฟ้มข้อมูลภาษาเอชทีเอ็มแอล เช่น แท็กคำสั่งที่ใช้ในการสร้างตารางคือแท็ก <Table> การสร้างแถวใช้คำสั่งแท็ก <TR> และการสร้างคอลัมน์โดยใช้คำสั่งแท็ก <TD> จากนั้นจะวิเคราะห์โครงสร้างภายในตารางซึ่งจะประกอบด้วย รูปภาพ เนื้อความ ภาพโฆษณาและภาพทั่วไปที่อยู่ภายใน ซึ่งจากหลักการเขียนภาษาเอชทีเอ็มแอล มีหลักการวิเคราะห์จากโครงสร้างดังแสดงในรูปที่ 1.1




**รูปที่ 1.2** ลักษณะเว็บไซต์และการวิเคราะห์ภาพจากภาษาเอชทีเอ็มแอล

แท็ก <A href> เป็นแท็กสำหรับใช้ในการเชื่อมโยง (Link) หน้าเว็บเพจทั้งหมดเข้าด้วยกัน[11] ซึ่งมีส่วนประกอบหลักอยู่ 2 ส่วนคือ ส่วนแรกตำแหน่งที่จะใช้เชื่อมโยงจากหน้าหนึ่งไปยังอีกหน้าหนึ่ง ซึ่งอาจเป็นข้อความหรือรูปภาพก็ได้ โดยตำแหน่งนี้จะเรียกว่าเอ็นเชอร์แท็ก (Anchor Tag) ซึ่งเมื่อคลิกตรงตำแหน่งนี้แล้วหน้าเว็บหลักก็จะทำการเชื่อมโยงไปยังหน้าเว็บเพจหรือเอกสารที่กำหนดไว้ทันที เพื่อจะไปดึงเอาแฟ้มข้อมูลที่เป็นเอชทีเอ็มแอลหรือเป็นภาพเคลื่อนไหวมาแสดงแทนหน้าเว็บเพจที่อยู่บนจอภาพขณะนั้น สำหรับส่วนที่สองเป็นส่วนของตำแหน่งที่กำหนดให้ไปหรือตำแหน่งที่มีข้อมูลอยู่ ซึ่งจะเขียนอยู่ในรูปของลักษณะการบอกตำแหน่งยูอาร์แอล (URL) และเส้นทาง (Path)

จากตารางรูปที่ 1.2 ส่วนที่อยู่ทางด้านขวาของภาพ (ส่วนA) จะแสดงโปรแกรมภาษาเอชทีเอ็มแอลของแฟ้มข้อมูลรูปภาพที่ใช้ในหน้าเว็บไซต์ที่อยู่ในตารางเดียวกัน เหมือนกันกับทางด้านขวาของภาพและส่วนที่อยู่ด้านซ้ายของภาพ (ส่วนB) แสดงแฟ้มข้อมูลที่เป็นรูปภาพ จากรูปเมื่อวิเคราะห์เป็นโครงสร้างต้นไม้ (Tree Structure) จะมีลักษณะดังนี้




**รูปที่ 1.3** การวิเคราะห์โดยใช้โครงสร้างต้นไม้

จากรูปตัวอย่าง ใช้สีแสดงสัญลักษณ์ สีขาวคือภาพโฆษณา (Advertisement) สีดำคือข้อความ (Content) และสีเทาคือไม่ชัดเจน (Unknown) โหนดแม่คือแท็กตาราง(<table>) ประกอบด้วยโหนดลูกคือแท็ก<td>ที่อยู่ในตารางเดียวกัน จากรูปซ้ายประกอบด้วยโหนดแม่ คือ A มีสมาชิกทั้งหมด 4 โหนดคือ A-1, A-2, A-3 และ A-4 เมื่อ A-1 เป็นสีขาวคือภาพโฆษณาแล้วอัลกอริทึมที่ใช้จะให้ผลว่าโหนดหรือแท็กอื่นที่อยู่ในตารางเดียวกัน แล้วมีความน่าจะเป็นโฆษณาเหมือนกัน เช่นกันกับรูปขวา โหนดแม่คือ B มีสมาชิกในโหนดเดียวกัน 1 โหนดคือ B-1 เมื่อ B-1 เป็นสีดำ คือเป็นข้อความ อัลกอริทึมที่ใช้จะให้ผลว่าโหนดอื่นที่อยู่ภายใต้ตารางเดียวกันมีความน่าจะเป็นข้อความเหมือนกัน แต่บางครั้งการวิเคราะห์โครงสร้างในลักษณะนี้อาจใช้ไม่ได้ผลนักกับเว็บไซต์ในปัจจุบัน เนื่องจากการทำโฆษณามีการปรับเปลี่ยนและการทำเว็บไซต์มีความหลากหลายในการสร้างรูปแบบมากขึ้น จึงต้องมีการปรับคุณลักษณะเด่นให้เป็นปัจจุบันตาม

สำหรับขั้นตอนการทำงานวิจัยนี้จะใช้ทั้งหมด 4 คุณลักษณะเด่น คือ ค่าอัตราส่วนของรูป (Ratio) คำที่อยู่ในแท็กรูปภาพ (Word) ลักษณะคำสั่งที่เป็นโฆษณา (Pop-coding) และลักษณะของรูปแบบการเขียนโปรแกรม (Type of file) โดยทำการวิเคราะห์จากโปรแกรมภาษาเอ็ชทีเอ็มแอล จากนั้นนำข้อมูลที่ได้มาทำการจำแนก (Classify) โดยใช้หลักการจำแนกแบบเบส์ (Bay Classifier) แล้วทำการทดสอบผลที่ได้โดยมีอัลกอริทึมแสดงขั้นตอนการทำงานของระบบ ซึ่งทำการประเมินผลโดยวัดจากค่าความถูกต้องเป็นร้อยละ(เปอร์เซ็นต์)และทำการวิเคราะห์และสรุปผลจากผลการทดลองงานวิจัยโดยแสดงเป็นตารางแสดงเมทริกซ์แสดงความสับสน (Confusion metrix) ของคุณลักษณะเด่นที่ใช้

สำหรับคุณลักษณะเด่น (Feature) ของความน่าจะเป็นของโครงสร้างต้นไม้ (HTML tree) จากงานวิจัยที่ผ่านมาจะพิจารณาจากโครงสร้างของหน้าเว็บไซต์เป็นหลัก โดยอ่านจากโปรแกรมภาษาเอ็ชทีเอ็มแอล ซึ่งในปัจจุบันการเขียนโปรแกรมบนเว็บไซต์ ซึ่งคำสั่งส่วนใหญ่ในที่ยังใช้อยู่เช่น <table> , <tr> , <td> และเริ่มมีการนำคำสั่ง  เข้ามาใช้งาน รวมถึงภาษาสไตล์

ชีต (Style sheet) ซึ่งเรียกว่า Cascading Style Sheets (CSS) เข้ามาใช้มากขึ้น เนื่องจากเพิ่มความเร็วในการโหลด ซึ่งทั้งหมดนี้สามารถนำมาทำเป็นโครงสร้างต้นไม้ได้ เพราะทุกเว็บไซต์มีลักษณะโครงสร้างที่สามารถเขียนเป็นโครงสร้างต้นไม้จากโปรแกรมเอ็ชทีเอ็มแอลเหมือนกัน

จากการศึกษาเกี่ยวกับภาพโฆษณาต่างๆ ส่วนใหญ่ได้มุ่งศึกษาทางอัลกอริทึมที่เกี่ยวกับการบล็อกภาพโฆษณาต่างๆ รวมถึงศึกษาโครงสร้างแฟ้มข้อมูลภาษาเอ็ชทีเอ็มแอลแล้วนำมาวิเคราะห์และปรับปรุงดังนี้

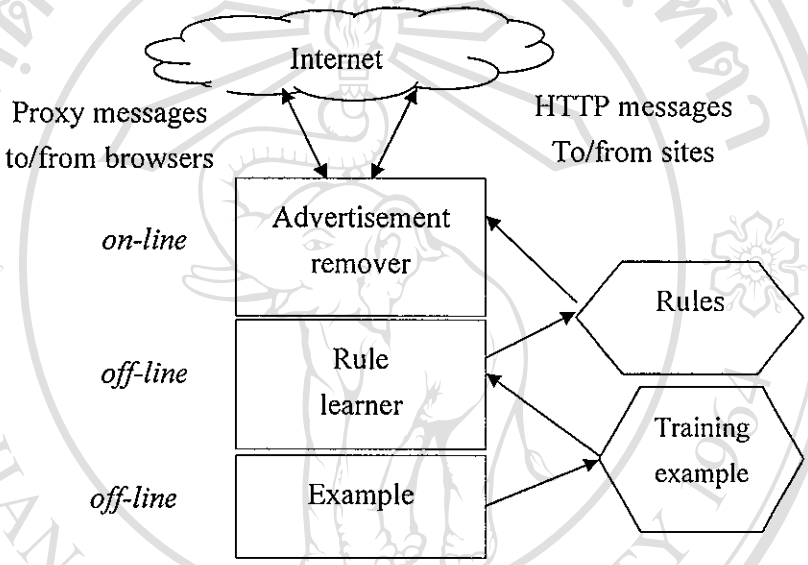
W.Cohen [5] ได้ศึกษาและพัฒนาเกี่ยวกับอัลกอริทึมที่ใช้ในลักษณะการเรียนรู้ (Rule Learning) และได้ใช้อัลกอริทึมนี้เปรียบเทียบกับกฎการเรียนรู้ เพื่อวัดประสิทธิภาพของอัลกอริทึม

M.Craven, D.Freitag, A.McCallum, T.Mitchell, K.Nigam and C.Quek. [6] ได้ทำการศึกษาเกี่ยวกับลักษณะการทำงานบนเว็บไซต์ เช่น การจัดและจำแนกประเภทข้อความ (Text classification) รวมถึงความสัมพันธ์ของการทำงานในแต่ละส่วนภายในเว็บไซต์ (Relation Learning) ซึ่งในการทดสอบข้อมูลได้มีการอ้างอิงยูอาร์แอลที่ใช้เชื่อมโยงไปยังเว็บไซต์หลัก (Primary pages) ของหน้าเว็บนั้น งานวิจัยนี้จะเรียกยูอาร์แอล ที่เป็นปัจจุบัน (Secondary pages) และเว็บไซต์หลักที่ได้ทำการเชื่อมโยง

H.Liu and L.Yu [8] ได้ศึกษาเกี่ยวกับการกลั่นกรองข้อมูลและข่าวสารจากการทำงานของภาษาเอ็ชทีเอ็มแอล โดยได้ศึกษาการทำงานผ่านเว็บไซต์โดยแฟ้มข้อมูลภาษาเอ็ชทีเอ็มแอลแล้วนำมาพัฒนาขึ้นมาโดยใช้ชื่อว่าเอสอาร์วี (SRV) ซึ่งเป็นอัลกอริทึมที่แสดงความสัมพันธ์ของข้อมูลที่ผ่านการกลั่นกรองและจัดหมวดหมู่แล้ว โดยในงานวิจัยนี้ได้แสดงให้เห็นหลักการทำงานของแฟ้มข้อมูลภาษาเอ็ชทีเอ็มแอลซึ่งสามารถนำมาศึกษาเกี่ยวกับงานวิจัยนี้ได้เช่นกัน

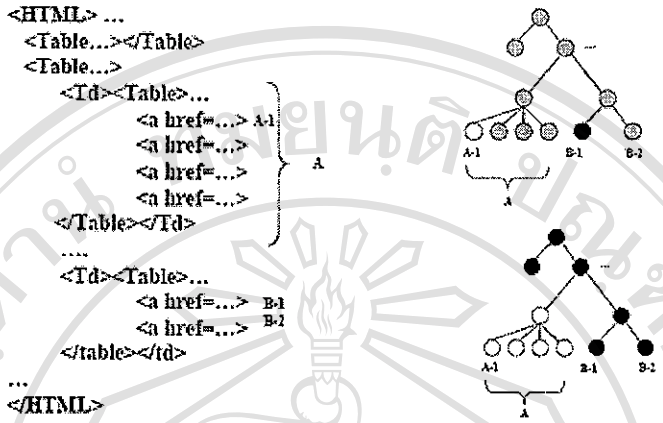
N. Kushmerick [9] ได้ศึกษาการนำภาพที่เป็นโฆษณาออกจากเว็บเพจโดยใช้ชื่อซอฟต์แวร์ในงานวิจัยว่า AdEater System ซึ่งใช้คุณลักษณะเด่นทั้งหมดคือ ขนาดของภาพซึ่งประกอบด้วย ความกว้าง (Width) , ความสูง (Height) , ค่าอัตราส่วนของภาพ (Aspect ratio) , ยูอาร์แอล (URL) , หัวข้อ (Caption) , จำนวนคำ (Text) หรือการใช้ข้อความที่เป็นการโฆษณา แล้วนำคุณลักษณะเด่นทั้งหมด 8 คุณลักษณะเด่น แล้วนำข้อมูลทั้งหมดมาทำการฝึกฝนข้อมูล (Train) แล้วทำการทดสอบข้อมูล (Test) โดยอัลกอริทึมที่ใช้ประกอบด้วย Ripper , ID3 ,Version Spaces จากนั้นได้ทำการพัฒนาและปรับปรุงโดยใช้กฎการเรียนรู้ (Machine Learning) แล้วใช้กฎ C4.5 ซึ่งเมื่อทำซอฟแวร์จากงานวิจัยนี้ชื่อว่า Ad Eater System สามารถใช้ได้ในการจำแนกภาพที่เป็นภาพโฆษณาและเป็นภาพอื่นที่ไม่ใช่ภาพโฆษณา ซึ่งจากปัญหาที่พบคือการเรียนรู้ของเครื่องวัดประสิทธิภาพได้ประมาณ 97% จากการเรียนรู้และการทดสอบโดยการฝึกฝน (Train) และการทดสอบ (Test) ซึ่งผลที่ได้มาจากข้อมูลในชุดเดียวกัน โดยแบ่งการฝึกและการทดสอบออกเป็น

10:90 ไปเรื่อยๆ จนครบ 100% แล้วแสดงผลที่สามารถทำการจำแนกได้ ซึ่งเมื่อทำโปรแกรมคอมพิวเตอร์จากงานวิจัยนี้ชื่อว่า AdEater System สามารถใช้ได้ในการจำแนกภาพที่เป็นภาพโฆษณาและเป็นภาพอื่นที่ไม่เป็นโฆษณาจากปัญหาที่พบคือการเรียนรู้ของเครื่องสามารถวัดประสิทธิภาพได้ประมาณ 97% จากการเรียนรู้และการทดสอบโดยใช้การฝึกฝนข้อมูลและการทดสอบข้อมูลในชุดเดียวกัน




**รูปที่ 1.4** ขั้นตอนการเรียนรู้โดยใช้ C4.5 Rule Learner

L.Shih and D.Karger [10] ได้ศึกษาการจำแนก (Classification) ภาพที่เป็นโฆษณาโดยใช้โครงสร้างต้นไม้ ซึ่งเป็นการศึกษาโครงสร้างของภาษาเอ็ชทีเอ็มแอลเป็นคุณลักษณะเด่นและใช้โครงสร้างยูอาร์แอล (URL) และลักษณะโครงสร้างของคำสั่งที่ใช้สร้างตารางในลักษณะโครงสร้างต้นไม้ (Table tree) จากนั้นจะหาค่าความน่าจะเป็นของโหนดที่เป็นโหนดแรกภายใต้โหนดแม่เดียวกัน ถ้าโหนดที่หนึ่งมีภาพที่เป็นโฆษณา แล้วโหนดที่อยู่ข้างๆ จะเริ่มเปลี่ยนไปตามการคำนวณจนท้ายที่สุดแล้วจะมีความน่าจะเป็นที่เป็นภาพโฆษณาทั้งกิ่ง เช่นเดียวกันกับลักษณะของโครงสร้างต้นไม้ จะเริ่มเปลี่ยนไปตามการคำนวณของโครงสร้างตารางไปเรื่อยๆ จนท้ายที่สุดแล้วจะมีความน่าจะเป็นที่เป็นภาพโฆษณาทั้งกิ่งเช่นเดียวกัน โดยนำค่าทั้งสองมาเป็นคุณลักษณะเด่น แล้วใช้ทฤษฎีในการจำแนกเปรียบเทียบกัน ตัวอย่างดังรูปที่ 1.5




รูปที่ 1.5 ขั้นตอนการวิเคราะห์โครงสร้างของภาษาเอ็ชทีเอ็มแอลตามโครงสร้างต้นไม้

จากนั้นจะวัดค่าความน่าจะเป็นของโฆษณาโดยดูจากโหนดที่มาจากกิ่ง (Tree) เดียวกัน แล้วใช้อัลกอริทึมแสดงผลเปรียบเทียบกันระหว่าง Naive Bayes และ Support Vector Machine (SVM) จากผลการทดลองพบว่าทั้งสองอัลกอริทึมที่ใช้ทำให้ได้สามารถบล็อกภาพที่เป็นโฆษณาได้และมีประสิทธิภาพที่แตกต่างกัน

W.Cohen [11] ได้ศึกษาลักษณะการใช้คำและเทคนิคการปรับปรุง (Update) เว็บไซต์โดยทั่วไปให้สามารถจัดจำแนกเป็นกลุ่มโดยใช้การเชื่อมโยงภายในเว็บไซต์และแสดงเทคนิคการใช้คำโดยมีการแยกคำเป็นกลุ่มที่มักใช้กันอย่างสม่ำเสมอเพื่อทำการเชื่อมโยงไปสู่ลักษณะหรือกลุ่มคำหลัก เช่น ถ้าเป็นถ้อยคำประเภทโฆษณาก็จะมีกลุ่มลักษณะหนึ่ง ถ้าเป็นข้อความประเภทข่าวก็จะมีลักษณะกลุ่มคำอีกประเภทหนึ่ง ซึ่งงานวิจัยนี้เรียกลักษณะอัลกอริทึมว่า Wrapper

David M.Blei, J.Andrew Bagnell and A.McCallum [12] ได้ศึกษาเกี่ยวกับความน่าจะเป็นสำหรับคำที่มีลักษณะใกล้เคียงกันเป็นรูปแบบคำ (Word format) ที่ใช้ในเว็บเพจทั่วไป ซึ่งจะมีรูปแบบที่แตกต่างกันในแต่ละเว็บเพจ ซึ่งงานวิจัยนี้พยายามหารูปแบบของคำที่ใช้ให้ได้เพื่อแยกลักษณะของคำว่ามีรูปแบบ (Formatting) และคำทั่วไป (Word content) ประโยชน์ของงานวิจัยนี้คือจะได้รู้รูปแบบคำที่เป็นโฆษณาหรือคำที่เป็นข้อความข่าวซึ่งใช้คุณลักษณะเด่นของคำช่วยในการตัดสินได้

William W.Cohen and Lee S.Jensen [13] ได้ศึกษาเกี่ยวกับโครงสร้างของอัลกอริทึม Wrapper ซึ่งเรียกว่า Wrapper-learning system โดยลักษณะของโครงสร้างคำและข้อความ (Structured documents) ในภาษาเอ็ชทีเอ็มแอลจะใช้วิธีการจัดความสัมพันธ์ของข้อมูลใน

คอมพิวเตอร์ (DOM:document object model) ซึ่งมีประโยชน์กับงานวิจัยนี้คือ สามารถแยกแท็กจากแฟ้มข้อมูลเอ็ชทีเอ็มแอลซึ่งเป็นเรื่องที่กำลังศึกษาในงานวิจัยนี้ว่าแท็กของภาษาเอ็ชทีเอ็มแอลนั้นมีประโยชน์ในการบอกโครงสร้าง (Structure) ในหน้าเว็บเพจได้

William W.Cohen , Matthew Hurst and Lee S. Jensen [14] ได้ศึกษาเกี่ยวกับการทำงานของเว็บไซต์ในปัจจุบันที่มีลักษณะเป็นฐานข้อมูล (Database) ขนาดใหญ่ เรียกว่า Wrapper หรือ WL2 รวมถึงลักษณะของ DOM-Level เป็นการแสดงหน้าเว็บเพจเพื่อเพิ่มความเร็วในการแสดงหน้าเว็บไซต์ทั่วไป มีประโยชน์เพื่อจะได้ทราบถึงขั้นตอนและวิธีการในการแสดงหน้าเว็บไซต์และการจัดระเบียบโครงสร้างการแสดงหน้าเว็บเมื่อเข้าสู่เว็บไซต์ที่ต้องการ

Justin Crites and Mathias Ricken [16] ได้ศึกษาเกี่ยวกับการบล็อกภาพโฆษณาเรียกว่า Mozilla Platform โดยในงานวิจัยนี้จะทำการบล็อกภาพโฆษณาบนเว็บบราวเซอร์ (Web browser) ประเภทบราวเซอร์ของไฟล์ฟ็อกซ์ (Firefox browser) โดยเป็นการคิดอัลกอริทึมใหม่ที่สามารถบล็อกจากแฟ้มข้อมูลภาษาเอ็ชทีเอ็มแอลโดยแบ่งเป็น

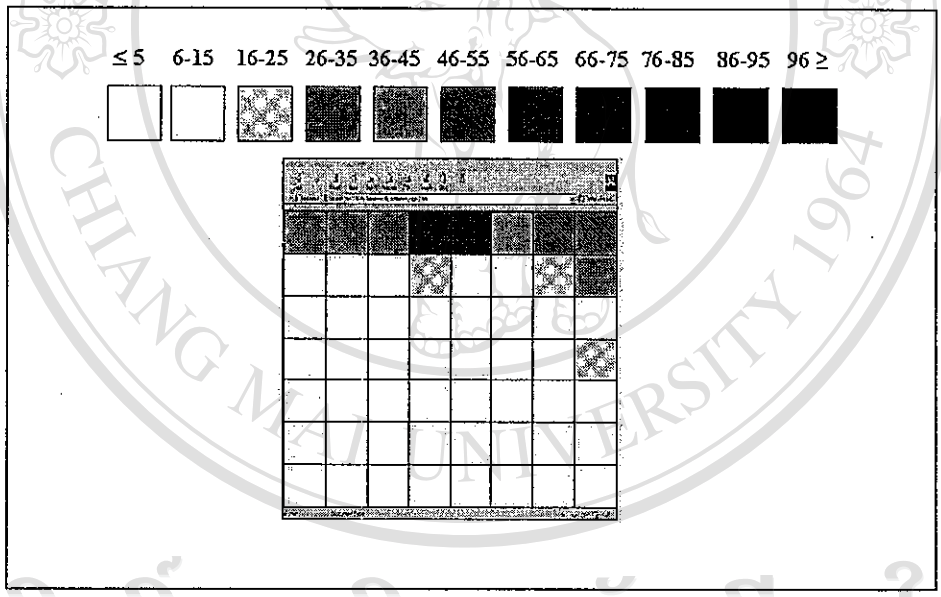
1. อัลกอริทึมที่สามารถแยกสัญลักษณ์จากยูอาร์แอลเรียกว่า Dynamic programming algorithm for LCS (Longest common subsequence)
2. เว็บอัพเดต (Blacklist web updates)
3. การบล็อกจากเส้นทาง (DOM path blocking) เป็นการบล็อกเมื่อเห็นแท็กที่น่าจะเป็นภาพโฆษณา เช่น <IMG> , <OBJECT> และ <IFRAME> เป็นต้น ซึ่งสำหรับงานวิจัยนี้เลือกทดสอบงานวิจัยโดยใช้เว็บบราวเซอร์เป็นไฟล์ฟ็อกซ์

X.Yin and W.Lee [17] ได้ศึกษาเกี่ยวกับวิธีการสำหรับการจำแนกในหน้าเว็บเพจออกเป็น 5 ส่วน คือ ส่วนข้อความ (Content) ส่วนการเชื่อมโยง (Related Links) ส่วนสำหรับแนะนำหรือให้ความช่วยเหลือ (Navigation and Support) ส่วนที่เป็นการโฆษณา (Advertisement) และส่วนของรูปแบบ (Form) โดยจะสร้างกราฟขึ้นมาในแต่ละส่วนและศึกษาการทำงานในแต่ละส่วน จากนั้นทำการทดสอบโดยใช้กฎการเรียนรู้ (Machine Learning) แล้วดูผลลัพธ์ ตรวจสอบความถูกต้องของงานวิจัย ซึ่งก็ให้ผลลัพธ์ที่ถูกต้องมากขึ้น

M. Kovacevic , M.Diligenti, M.Gori and V.Milutinovic [18] ได้ศึกษาเกี่ยวกับการจำแนกภายในหน้าเว็บไซต์ตามโครงสร้างของหน้าเว็บ โดยจะแบ่งออกเป็นส่วนต่างๆทั้งหมด 6 ส่วนโดยอ่านจากโปรแกรมภาษาเอ็ชทีเอ็มแอล ซึ่งทั้ง 6 ส่วนประกอบด้วย ส่วนหัว (Head) ส่วนหัวข้อ (Title) ส่วนกลาง (Centre) ส่วนขอบซ้าย (Left menu) ส่วนขอบขวา (Right menu) และส่วนล่าง (Footer) จากนั้นใช้ทฤษฎีของเบส์ (Naive Bayes) เป็นตัวแยก(Classifier) ซึ่งจากผลที่ได้แสดงให้เห็นว่ามีความถูกต้องมากขึ้น

C.Lee, M.Kan and S.Lai [19] ได้ศึกษาและนำเสนอ Parcels เป็นลักษณะการจำแนกอีกประเภทหนึ่งซึ่งสามารถแสดงให้เห็นลักษณะโครงสร้างตารางของหน้าเว็บไซต์ โดยผลลัพธ์ที่ได้จากงานวิจัยนี้สามารถทำให้มีความผิดพลาด (Error) น้อยลง

เว็บไซต์ http://www.design.eti.br [20] เป็นงานวิจัยที่แสดงถึงตำแหน่งของโฆษณาประเภทแบนเนอร์ในตำแหน่งต่างๆ ภายในเว็บไซต์ ซึ่งตำแหน่งแต่ละส่วนจะมีผลต่อการโฆษณา เนื่องจากการโฆษณาต้องวิเคราะห์ว่าตำแหน่งส่วนไหนในหน้าเว็บจะได้รับความสนใจในการเข้าชม ดังนั้นจึงได้มีการวิเคราะห์และสรุปผลการแสดงตำแหน่งของภาพโฆษณาแบนเนอร์ที่มีอยู่ทั่วไปในเว็บไซต์ ซึ่งมีแนวคิดคล้ายกับงานวิจัยนี้โดยใช้ตำแหน่งภายในหน้าเว็บไซต์มาช่วยในการวิเคราะห์ข้อมูลภาพโฆษณาประเภทแบนเนอร์ได้มากยิ่งขึ้น รูปที่ 1.6 แสดงตัวอย่างการวางตำแหน่งของโฆษณาแบนเนอร์ในหน้าเว็บไซต์




รูปที่ 1.6 การวิเคราะห์ลักษณะการวางตำแหน่งของโฆษณาแบนเนอร์