

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงรายละเอียดของทฤษฎีที่เกี่ยวข้องและเทคนิควิธีการต่างๆ ที่นำมาประยุกต์ใช้ในงานวิจัย ซึ่งมีรายละเอียดดังต่อไปนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้องกับงานวิจัยฉบับนี้ประกอบด้วย มาตรฐานเว็บ (Web Standards), มาตรฐานเว็บไซต์ที่ทุกคนเข้าถึงได้ (WCAG: Web Content Accessibility Guidelines), หัวเรื่อง (Heading), วิไอพีเอสอัลกอริทึม (VIPS algorithms), ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) โดยมีรายละเอียดดังต่อไปนี้

2.1.1 มาตรฐานเว็บ

บทนี้จะกล่าวถึงทฤษฎีที่เกี่ยวข้องกับมาตรฐานเว็บที่ใช้ในการวิจัยนี้ซึ่งประกอบไปด้วย เอกซ์ทีเอ็มแอล (HTML), สไคชีท (CSS: Cascading Style Sheets), ดอม (DOM: Document Object Model)

2.1.1.1 เอกซ์ทีเอ็มแอล

เอกซ์ทีเอ็มแอล หรือ Hypertext Markup Language เป็นภาษาคอมพิวเตอร์รูปแบบหนึ่งที่มีโครงสร้างการเขียนโดยอาศัยป้ายระบุ (Tag) ควบคุมรูปแบบของการแสดงผลของข้อความ, รูปภาพ หรือวัตถุอื่นๆ ผ่านโปรแกรมเบราว์เซอร์ แต่ละป้ายระบุอาจจะมีส่วนขยายที่เรียกว่าแอตทริบิวต์ (Attribute) สำหรับระบุหรือควบคุมการแสดงผลของเว็บได้ด้วย โดยทั่วไปเอกสารเอกซ์ทีเอ็มแอล ประกอบด้วยส่วนป้ายระบุ <Head> และป้ายระบุ <Body> ของเอกสาร

ป้ายระบุจะเขียนโดยกำกับด้วยเครื่องหมาย “<” และ “>” โดยที่แท็กจะเป็นตัวกำหนดรูปแบบของการแสดงผลหรือโครงสร้างการแสดงผลเอกสารนั้น ข้อความที่ต้องการกำหนดรูปแบบของการแสดงผลจะถูกกำกับระหว่างป้ายระบุเปิด <...> และป้ายระบุปิด </...> เช่น <H1>สวัสดี</H1>

เป็นการกำหนดรูปแบบของการแสดงผลของข้อความ “สวัสดี” ด้วยตัวอักษรขนาดใหญ่ ตัวอย่างของเอกสาร HTML อย่างง่ายแสดงในภาพที่ 2.1

ภาพที่ 2.1
ตัวอย่างเอกสาร HTML

```
<HTML>
  <HEAD>
    <TITLE>ตัวอย่างเว็บอย่างง่าย</TITLE>
  </HEAD>
  <BODY>
    <H1>สวัสดี</H1>
  </BODY>
</HTML>
```

นอกจากป้ายระบุเลขที่เอมแอลแล้วยังมีแอตทริบิวต์ที่เป็นส่วนขยายความสามารถของป้ายระบุจะต้องใส่ภายในเครื่องหมาย < > ในส่วนป้ายระบุเปิดเท่านั้น ป้ายระบุคำสั่งเลขที่เอมแอลแต่ละคำสั่งจะมีแอตทริบิวต์แตกต่างกันไป และมีจำนวนไม่เท่ากัน การระบุแอตทริบิวต์มากกว่า 1 แอตทริบิวต์ ให้ใช้ช่องว่างเป็นตัวคั่น เช่น แอตทริบิวต์ของป้ายระบุ <H1> สามารถเขียนได้โดยภาพที่ 2.2

ภาพที่ 2.2
ตัวอย่างการใช้แอตทริบิวต์ของป้ายระบุเลขที่เอมแอล

```
<H1 ALIGN="LEFT">สวัสดี</H1>
```

2.1.1.2 สไตล์ชีท

สไตล์ชีท (CSS) เป็นการกำหนดรูปแบบของเอกสารเอกซ์ทีเอ็มแอลหรือเว็บเพจ เช่น การจัดตำแหน่งของข้อความกับรูปภาพให้ตรงตามความต้องการของผู้ออกแบบเว็บเพจ และสไตล์ชีทสามารถแยกส่วนเนื้อหาของเอกสารกับส่วนกำหนดรูปแบบของการแสดงผลของเอกสารออกจากกันได้ซึ่งช่วยลดความซับซ้อนของส่วนเนื้อหาของเอกสารเอกซ์ทีเอ็มแอลได้ โครงสร้างของสไตล์ชีทอยู่ภายใต้ไวยากรณ์ <Style> โดยประกอบด้วยรายการของกฎหลายอัน ภายในกฎแต่ละอันจะแบ่งออกเป็นส่วนของชื่อและรูปแบบของการแสดงผลซึ่งภายในส่วนแสดงผลยังแบ่งออกเป็นสองส่วนด้วยเครื่องหมายโคลอน (:) ส่วนที่อยู่ด้านหน้าเครื่องหมายโคลอนจะหมายถึงคุณสมบัติ (Property) ที่ต้องการกำหนดรูปแบบ และส่วนที่อยู่หลังเครื่องหมายโคลอนจะเป็นค่าที่ต้องการกำหนดให้คุณสมบัติ ตัวอย่างของสไตล์ชีทสามารถแสดงได้ดังภาพที่ 2.3

ภาพที่ 2.3

ตัวอย่างการประกาศใช้สไตล์ชีท

```
<STYLE>
  P{COLOR: RED;}
  .Heading{MARGIN: 1em 0px}
  #lin1{FONT-WEIGHT: bold; FONT-SIZE: 50%}
</STYLE>
```

เมื่อต้องการกำหนดรูปแบบของการแสดงผลให้กับข้อความของเอกสารเอกซ์ทีเอ็มแอลก็สามารถทำได้ด้วยการเรียกใช้ชื่อกฎที่ตั้งขึ้นในส่วนเนื้อหาดังภาพที่ 2.4 เป็นการเรียกใช้สไตล์ชีทที่กำหนดไว้ในภาพที่ 2.3 โดยส่วนข้อความที่ถูกครอบด้วยไวยากรณ์ <P class=Heading > จะถูกกำหนดให้คุณสมบัติ Margin มีค่าเท่ากับ 1em 0px ซึ่งในกรณีที่มีความต้องการกำหนดรูปแบบของการแสดงผลให้มีคุณสมบัติที่มี Margin ค่าเดียวกันก็สามารถเรียกสไตล์ชีทที่กำหนดไว้เช่นเดียวกัน และเมื่อมีต้องการเปลี่ยนรูปแบบของการแสดงผลของข้อความก็สามารถแก้ไขได้จากที่เดียวกัน

ภาพที่ 2.4
ตัวอย่างการเรียกใช้สไตร์ชีท

```
<P CLASS=Heading>
    <H1>สวัสดี</H1>
    <span>สวัสดี เมืองไทย</span>
</P>
```

2.1.1.3 ดอม

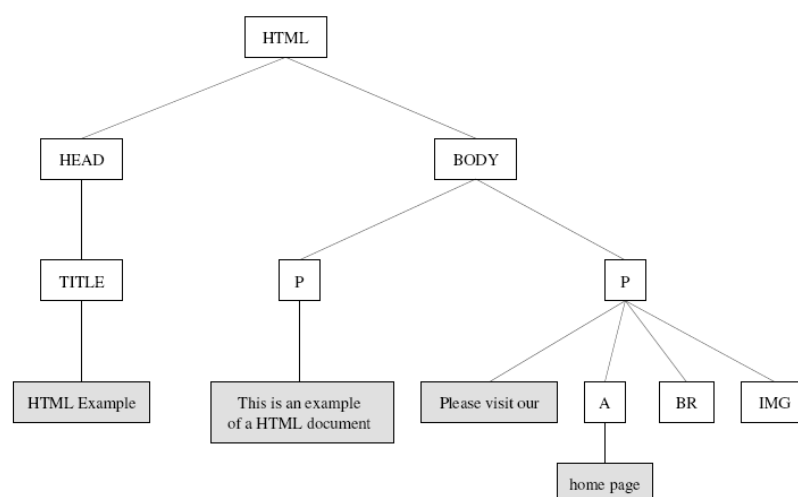
ดอม (DOM) เป็นแพลตฟอร์มและภาษาสื่อกลางระหว่างโปรแกรมกับสคริปต์ต่างๆ สามารถเข้าถึงและปรับปรุง เนื้อหา, โครงสร้าง และสไตร์ชีทของเอกสารเอกซีเอ็มแอล ดอมจะมององค์ประกอบทั้งหมดของเว็บเพจเป็นโหนดต้นไม้ (Tree node) ซึ่งสามารถแบ่งประเภทของโหนดได้ 3 ชนิดได้แก่

1. โหนดสมาชิก (Element node) คือ ป้ายระบุของเอกซีเอ็มแอลเช่น ป้ายระบุ <head>, <P>, <H1> เป็นต้น
2. โหนดข้อความ (Text node) คือ ส่วนที่เป็นข้อความของแต่ละโหนดสมาชิก หรือส่วนที่เป็นข้อความของป้ายระบุเอกซีเอ็มแอล
3. โหนดแอตทริบิวต์ (Attribute node) คือ แอตทริบิวต์ของป้ายระบุต่างๆ ของเอกสารเอกซีเอ็มแอล เช่น title , style เป็นต้น

โหนดสมาชิกอาจประกอบด้วยโหนดลูก (Child node) จำนวนเท่าใดก็ได้ซึ่งสามารถเป็นได้ทั้งโหนดข้อความหรือโหนดสมาชิกอื่นๆ แต่โหนดข้อความไม่สามารถมีโหนดลูกได้ ภาพที่ 2.5 แสดงตัวอย่างของการนำเสนอเอกสารเอกซีเอ็มแอลด้วยดอม กล้องสี่เหลี่ยมที่มีสีเทาแสดงถึงโหนดข้อความ ส่วนกล้องสี่เหลี่ยมที่มีสีขาวแสดงถึงโหนดสมาชิก

ภาพที่ 2.5

ตัวอย่างการแสดงผลเอกสารเอชทีเอ็มแอลด้วยคอม



2.1.2 มาตรฐานเว็บไซต์ที่ทุกคนเข้าถึงได้

เว็บไซต์ที่ทุกคนเข้าถึงได้(Web Accessibility) คือ เว็บไซต์ที่สามารถให้บริการข้อมูลและข่าวสารกับทุกคนอย่างเท่าเทียมกัน โดยคำนึงถึงอุปสรรคในการรับรู้ข่าวสาร เช่น ความพิการทางด้านการเห็น การได้ยิน เป็นต้น โดยมีการกำหนดแนวทางมาตรฐานในการออกแบบเว็บไซต์ที่ทุกคนสามารถเข้าถึงได้เพื่อแก้ปัญหาของการเข้าถึงข้อมูล เช่นคนพิการทางด้านการมองเห็นไม่สามารถอ่านข้อมูลด้วยประสาทตาได้ จึงต้องมีทางเลือกอื่นสำหรับผู้พิการกลุ่มนี้ ซึ่งอาจจะเปลี่ยนข้อมูลในรูปตัวอักษรให้เป็นข้อมูลเสียงแล้วใช้อุปกรณ์ช่วยเหลือพิเศษในการอ่านหน้าจอบนการแสดงผลของบราวเซอร์ ณ ขณะนั้น

องค์กร W3C จึงได้กำหนดมาตรฐานเว็บไซต์ที่ทุกคนสามารถเข้าถึงได้ ปัจจุบันอยู่ในเวอร์ชัน WCAG 2.0 (WCAG) ที่ประกาศใช้ในปี 2007 ซึ่งมีโครงสร้างดังนี้

หลักการ (Principle) มี 4 หลักการตามวัตถุประสงค์ในการออกแบบเว็บไซต์

- หลักการที่ 1 – ผู้อ่านต้องสามารถรับรู้เนื้อหาได้ ซึ่งแบ่งเป็นแนวทางย่อยดังต่อไปนี้
 - แนวทางที่ 1.1 - จัดเตรียมข้อมูลที่เป็นข้อความ (Text) แทน เนื้อหาที่มีรูปแบบเป็นอื่น

- แนวทางที่ 1.2 - จัดเตรียมข้อความบรรยายที่ตรงกับเหตุการณ์ในสื่อมัลติมีเดีย
- แนวทางที่ 1.3 - การออกแบบโครงสร้าง และเนื้อหา ต้องสามารถทำงานเป็นอิสระจากกันและกัน
- แนวทางที่ 1.4 - ต้องมั่นใจได้ว่าพื้นหน้าและพื้นหลัง(สีและเสียง) ต้องมีความแตกต่างกันมากพอที่ผู้ใช้จะสามารถแยกแยะได้
- หลักการที่ 2 - องค์ประกอบต่าง ๆ ของการอินเตอร์เฟสกับเนื้อหาจะต้องใช้งานได้ ซึ่งแบ่งเป็นแนวทางย่อย ดังต่อไปนี้
 - แนวทางที่ 2.1 - การทำงานทุกอย่างต้องรองรับการใช้งานจากคีย์บอร์ดได้
 - แนวทางที่ 2.2 - จัดเตรียมเวลาให้เพียงพอในการอ่าน หรือการกระทำใดๆ ของข้อมูล สำหรับผู้ใช้ที่เป็นคนพิการ
 - แนวทางที่ 2.3 - ไม่สร้างเนื้อหาที่อาจเป็นสาเหตุของอาการลมชัก
 - แนวทางที่ 2.4 - จัดเตรียมทางช่วยเหลือสำหรับผู้ใช้ในการสืบค้นเนื้อหา รู้ว่าตัวเองอยู่ในตำแหน่งใดในเนื้อหา และท่องไปในเนื้อหานั้นได้
- หลักการที่ 3 - ผู้ใช้สามารถเข้าใจเนื้อหา และส่วนควบคุมการทำงานต่างๆ ได้
 - แนวทางที่ 3.1 - สร้างเนื้อหาให้สามารถอ่านและเข้าใจได้
 - แนวทางที่ 3.2 - การทำงานของระบบต่างๆ หรือการแสดงผลบนหน้าเว็บ ต้องเป็นสิ่งที่ผู้ใช้สามารถคาดเดาได้
 - แนวทางที่ 3.3 - จัดเตรียมส่วนการช่วยเหลือให้ผู้ใช้ให้สามารถกรอกข้อมูลได้ถูกต้อง
- หลักการที่ 4 - เนื้อหาต้องมีความยืดหยุ่นที่จะทำงานกับเทคโนโลยีเว็บในปัจจุบันและอนาคตได้ (รวมถึงเทคโนโลยีสิ่งอำนวยความสะดวก)

- แนวทางที่ 4.1 – รองรับการใช้งานร่วมกับตัวแทนผู้ใช้ (User Agent) ได้ทั้งในปัจจุบัน และอนาคต (รวมถึงเทคโนโลยีสิ่งอำนวยความสะดวก)

เกณฑ์ความสำเร็จ (Success Criteria) เป็นตัวบอกถึงระดับความสำเร็จของหัวข้อแนวทางที่จะทำให้เป็นไปตามหลักการ ซึ่งสามารถแบ่งออกเป็น 3 ระดับ

- เกณฑ์ระดับ A เป็นเกณฑ์ที่ผู้พัฒนาเว็บไซต์ “ต้อง” ปฏิบัติตามเพื่อให้คนพิการเข้าถึงเว็บไซต์ได้
- เกณฑ์ระดับ AA เป็นเกณฑ์ระดับสำคัญรองลงมาที่ผู้พัฒนาเว็บไซต์ “ควรจะ” ปฏิบัติตามเพื่อให้คนพิการเข้าถึงเว็บไซต์ได้ง่ายขึ้น
- เกณฑ์ระดับ AAA เป็นเกณฑ์ที่ผู้พัฒนาเว็บไซต์ “อาจจะ” ปฏิบัติตามเพื่อให้คนพิการเข้าถึงเว็บไซต์ได้ง่ายที่สุด

2.1.3 หัวเรื่อง

เอกสารเชททีเอ็มแอลส่วนใหญ่จะมีการบอกหัวเรื่องในแต่ละส่วนย่อย เพื่อบอกให้ผู้ใช้รู้ว่ารายละเอียดที่จะกล่าวถึงมีเนื้อหาเกี่ยวกับอะไร หัวเรื่องใช้ป้ายระบุ <H1>~<H6> ในการกำหนดรูปแบบของการแสดงผล ซึ่งถูกแบ่งออกเป็น 6 ระดับตามความสำคัญ ซึ่ง ป้ายระบุ <H1> จะมีระดับความสำคัญมากที่สุดและ ป้ายระบุ <H6> จะมีระดับความสำคัญน้อยที่สุด (Heading) และการใช้ป้ายระบุ <H1>~<H6> ไม่ควรที่จะข้ามระดับของการใช้ เช่น ใช้ป้ายระบุ <H1> กับ ป้ายระบุ <H3> โดยไม่มีการใช้ป้ายระบุ <H2> ผู้มีสายตาปกติสามารถใช้หัวเรื่องเป็นเครื่องมือสำหรับการแบ่งแยกระหว่างเนื้อหาต่างๆ ผู้พิการทางสายตาก็สามารถใช้ประโยชน์จากหัวเรื่องได้เช่นกันถ้าหัวเรื่องนั้นใช้ป้ายระบุหัวเรื่องในการกำหนดรูปแบบของการแสดงผล ตัวอย่างของหัวเรื่องที่ใช้ป้ายระบุหัวเรื่องในการกำหนดรูปแบบของการแสดงผลดังภาพที่ 2.6

ภาพที่ 2.6

ตัวอย่างหัวเรื่องที่ใช้ป้ายระบุหัวเรื่องในการกำหนดรูปแบบของการแสดงผล (Jim, 2006)



ป้ายระบุหัวเรื่องในเอชทีเอ็มแอลเป็นหนึ่งในป้ายระบุที่มีความสำคัญมากที่สุดในการแสดงถึงโครงสร้างของเว็บเพจซึ่ง WCAG 2.0 นำเสนอแนวทางการในออกแบบเว็บไซต์ที่ทุกคนสามารถเข้าถึงได้แต่ละแนวทางจะมีเทคนิคสำหรับการออกแบบเว็บไซต์ให้เป็นไปตามแนวทางนั้น (WCAG) ในงานวิจัยนี้จะกล่าวถึงเพียงแค่แนวทางการใช้ป้ายระบุหัวเรื่องในการสร้างเว็บไซต์ที่ทุกคนเข้าถึงได้ (รายละเอียดฉบับสมบูรณ์ของแนวทางการออกแบบเว็บไซต์ให้ทุกคนเข้าถึงได้สามารถดูได้ที่ <http://www.w3.org/TR/WCAG20> ดังนี้

- แนวทางที่ 1.3.1 ข้อมูล, โครงสร้าง, และความสัมพันธ์กัน ซึ่งสื่อความหมายผ่านการนำเสนอสามารถคงรูปแบบการนำเสนอแม้ว่าจะถูกแสดงผลในเทคโนโลยีช่วยเหลืออื่นๆ หรืออยู่ในรูปของข้อความ
 - เทคนิคเอชที 42 : ใช้ป้ายระบุหัวเรื่องในการกำหนดรูปแบบของการแสดงผลของหัวเรื่อง แสดงตัวอย่างดังภาพที่ 2.7 ในการใช้ป้ายระบุหัวเรื่องในการกำหนดรูปแบบของการแสดงผลของหัวเรื่องในการระบุว่าเป็นเมนูหรือส่วนใดเป็นเนื้อหาหลัก

ภาพที่ 2.7

การใช้ป้ายระบุหัวเรื่องในการกำหนดรูปแบบของการแสดงผลของหัวเรื่อง (WCAG)

```
<!-- Logo, banner graphic, search form, etc. -->
<h2>Navigation</h2>
<ul>
  <li>
    <a href="about.htm">About us</a>
  </li>
  <li>
    <a href="contact.htm">Contact us</a>
  </li>
  ...
</ul>
<h2>All about headings</h2>
<!-- Text, images, other material making up the main content... -->
```

- แนวทางที่ 2.4.1 เตรียมวิธีการสำหรับการข้ามผ่านส่วนของเนื้อหาที่ถูกแสดงผลซ้ำกันในหลายหน้าเว็บเพจ
 - เทคนิคเลขที่ 69: เตรียมป้ายระบุหัวเรื่องที่เริ่มต้นของแต่ละส่วนย่อยของเนื้อหา แสดงตัวอย่างดังภาพที่ 2.8 ในการใช้ป้ายระบุหัวเรื่องในการเริ่มต้นของแต่ละส่วนย่อยของหน้าค้นหาโดยแต่ละหัวเรื่องของส่วนย่อยใช้ป้ายระบุ <h2> ในการกำหนดรูปแบบของการแสดงผล

ภาพที่ 2.8

การใช้ป้ายระบุหัวเรื่องที่เริ่มต้นส่วนย่อยของแต่ละเนื้อหา (WCAG)

```
<h1>Search Technical Periodicals</h1>

  <h2>Search</h2>

  <form action="search.php">

    <p>

      <label for="searchInput">Enter search topic:

</label>

      <input type="text" size="30" id="searchInput">

      <input type="submit" value="Go">

    </p>

  </form>

  <h2>Available Periodicals</h2>

  <div class="jlinks">

    <a href="pcoder.com">Professional Coder</a> |

    <a href="algo.com">Algorithms</a> |

    <a href="jse.com">Journal of Software Engineering</a>

  </div>

  <h2>Search Results</h2>

  ... search results are returned in this section ...
```

- แนวทางที่ 2.4.6 ใช้หัวเรื่องและฉลากในการบรรยายหัวข้อหรือวัตถุประสงค์
 - เทคนิคที่ 130: เตรียมหัวเรื่องที่สื่อความหมาย แสดงตัวอย่างดังภาพที่ 2.9 ในการใช้หัวเรื่องที่สื่อความหมายและใช้ลำดับของป้ายระบุหัวเรื่องในการแยกหัวเรื่องย่อย

ภาพที่ 2.9
การใช้หัวเรื่องที่สื่อความหมาย (WCAG)

<h1>Disaster preparation</h1>

<h2>Flood preparation</h2>

<h2>Fire preparation</h2>

- แนวทางที่ 2.4.10 : เตรียมหัวเรื่องสำหรับส่วนย่อยของเว็บเพจเมื่อเว็บเพจนั้นแบ่งออกเป็นส่วนย่อย
 - เทคนิคที่ 141: จัดการเว็บเพจโดยเตรียมหัวเรื่องในการบรรยายส่วนย่อยที่จะกล่าวถึง แสดงตัวอย่างดังภาพที่ 2.10 ในการใช้หัวเรื่องบรรยายส่วนย่อยที่จะกล่าวถึง

ภาพที่ 2.10
การใช้หัวเรื่องบรรยายส่วนย่อยที่จะกล่าวถึง (WCAG)

```

<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>Cooking techniques</title>
  </head>
  <body>
    <h1>Cooking techniques</h1>
    ... some text here ...

    <h2>Cooking with oil</h2>
    ... text of the section ...
    <h3>Sautéeing</h3>
    ....
    <h3>Deep frying</h3>

    <h2>Cooking with butter</h2>
    .. text of the section ...

  </body>
</html>

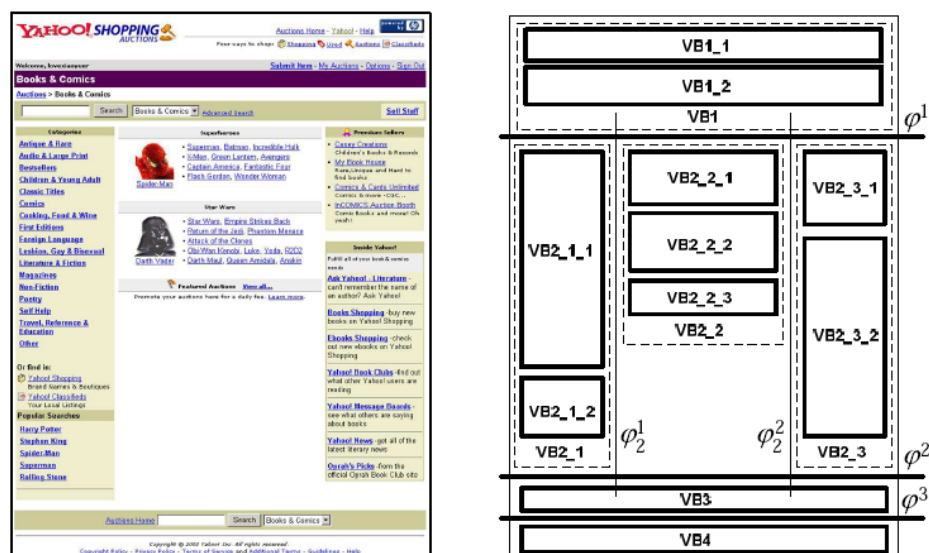
```

2.1.4 วิไอพีเอสอัลกอริทึม

วิไอพีเอสอัลกอริทึม (VIPS : Vision based Pages Segmentation) (Deng, 2003) เป็นอัลกอริทึมที่ใช้ในการแบ่งหน้าเว็บเพจออกเป็นส่วนย่อยซึ่งมาจากการรวมโครงสร้างแบบดอม (DOM Structure) และ โครงสร้างเนื้อหา (Vision-based Content Structure) จุดประสงค์หลักของวิไอพีเอสอัลกอริทึมคือ เพื่อแบ่งหน้าเว็บเพจให้มีลักษณะตามโครงสร้างของเนื้อหาสำคัญ โดยอยู่บนพื้นฐานของการแบ่งเนื้อหาที่สำคัญตามความหมาย โครงสร้างของวิไอพีเอสอัลกอริทึมแสดงได้ดังภาพที่ 2.11

ภาพที่ 2.11

โครงสร้างของเว็บเพจที่แบ่งเป็นส่วนย่อยด้วยไวโอไฟเอสอัลกอริทึม (Deng, 2003)

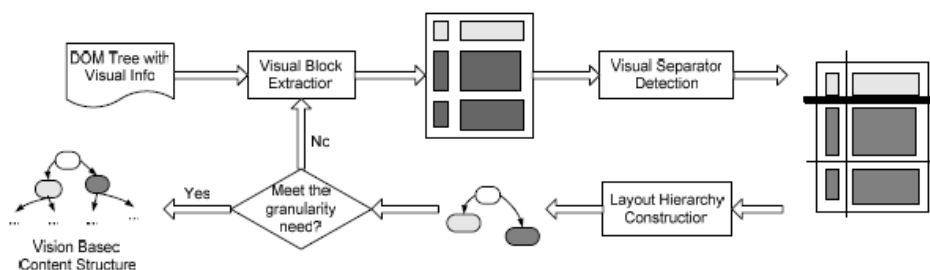


ขั้นตอนในการแบ่งเว็บเพจออกเป็นส่วนย่อยนั้นสามารถแสดงได้ดังภาพที่ 2.12 แบ่งเป็นขั้นตอนดังนี้

- 1) การตัดออกเป็นส่วนย่อย (Visual Segmentation Extraction) ขั้นตอนนี้จะทำการตัดหน้าเว็บเพจออกเป็นส่วนย่อยโดยที่แต่ละส่วนย่อยมีเนื้อหาคล้ายกัน
- 2) การตรวจหาตัวแบ่งแยก (Visual Separator Detection) ขั้นตอนนี้จะเป็นการกำหนดส่วนที่ถูกแบ่งออกเป็นส่วนย่อยให้แยกออกจากกันโดยตัวแบ่งแยกซึ่งอาจจะแบ่งตาม แนวตั้งหรือแนวนอนก็ได้ แล้วให้ค่าน้ำหนักของแต่ละส่วนตามความสำคัญ
- 3) การสร้างโครงสร้างของเนื้อหา (Content Structure Construction) ในขั้นตอนนี้จะทำการสร้างส่วนของเนื้อหาที่สำคัญให้มีลักษณะเป็นโครงสร้างที่ชัดเจน
- 4) ทำกระบวนการข้างต้นซ้ำอีก (Iterating the Above Step) ขั้นตอนนี้เป็นขั้นตอนสุดท้ายของการทำงานของไวโอไฟเอสอัลกอริทึม ซึ่งจะทำตามข้อ 1-3 ที่กล่าวมาข้างต้นซ้ำจนกว่าจะสามารถพบส่วนตรงกับความต้องการของผู้ใช้

ภาพที่ 2.12

ขั้นตอนการทำงานของวิไอพีเอสอัลกอริทึม (Deng, 2003)

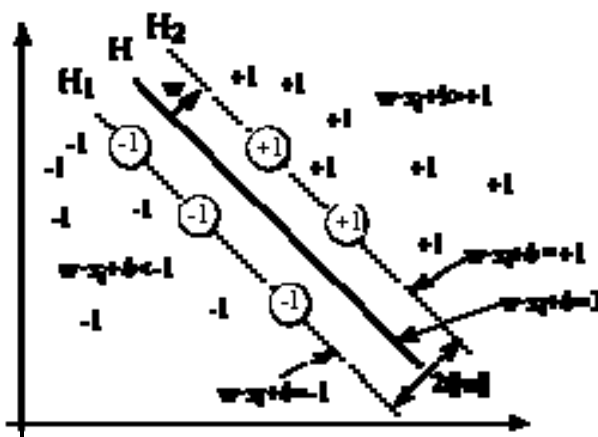


2.1.5 ซัพพอร์ตเวกเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีน (Cortes, 1995; Vapnik, 1998) เป็นเทคนิคที่ใช้ในการจำแนกค่าคุณลักษณะของสองกลุ่มข้อมูลที่วางอยู่ในพื้นที่คุณลักษณะ (Feature space) ออกจากกันโดยสร้างเส้นแบ่ง (Plane) ซึ่งอาศัยหลักการของการหาสมบัติของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ เพื่อค้นหาเส้นแบ่งที่แยกแยะกลุ่มข้อมูลได้ดีที่สุด (Optimal Separating Hyperplane)

ภาพที่ 2.13

แสดงเส้นแบ่งแยกกลุ่มข้อมูลที่ดีที่สุด



จากภาพที่ 2.13 แสดงเส้นแบ่งแยกข้อมูลที่ดีที่สุด เส้นทึบเป็นเส้นแบ่งแยกข้อมูลและเส้นประ 2 เส้นที่ขนานกับเส้นทึบจะเป็นเส้นขอบ (Margin) ที่บอกถึงขอบเขตที่เส้นทึบสามารถเคลื่อนที่ได้ภายในขอบเขตนี้โดยไม่ทำให้เกิดการจำแนกที่ผิดพลาด ซัพพอร์ตเวกเตอร์แมชชีนจะค้นหาเส้นทึบที่มีระยะเพิ่มเส้นขอบที่มากที่สุดซึ่งให้ผลลัพธ์ที่ดีที่สุดในการจำแนกกลุ่มข้อมูลด้วยความถูกต้องในการจำแนกมากที่สุด (Maximum accuracy) และมีค่าความผิดพลาดน้อยที่สุด (Minimum error rate)

จุดมุ่งหมายของซัพพอร์ตเวกเตอร์แมชชีนคือค้นหาค่าของระยะห่างที่มากที่สุดของ 2 เส้นประที่ขนานซึ่งสามารถแสดงได้ดังสมการที่ 2.1 และ 2.2 โดยที่ระยะห่างระหว่าง 2 เส้นประเท่ากับ $2/\|w\|$

$$x \cdot w + b = 1 \quad (2.1)$$

$$x \cdot w + b = -1 \quad (2.2)$$

ในการหาค่าสูงสุดของเส้นขนาด $M = 2/\|w\|$ ควรที่จะให้ค่า $\|w\|$ น้อยที่สุดภายใต้ข้อจำกัดของสมการที่ 2.1 และ 2.2 และสามารถเขียนสรุปได้ดังสมการที่ 2.3, 2.4 และ 2.5

$$x_i \cdot w + b \geq +1 \text{ กำหนด } y_i = +1 \quad (2.3)$$

$$x_i \cdot w + b \leq -1 \text{ กำหนด } y_i = -1 \quad (2.4)$$

$$y_i(x_i \cdot w + b) - 1 \geq 0 \quad (2.5)$$

เมื่อ x คือ ค่าลักษณะเด่น

w คือ ค่าความชัน

b คือ ค่าคงที่ (ค่าตัดแกน y)

x_i เป็น เวกเตอร์คุณลักษณะ (Feature vector) ของชุดการสอน i^{th} ซึ่งแทน n เวกเตอร์มิติ (dimensional vector) และ y_i แทนคลาสฉลาก (Label Class : +1 หรือ -1) ของชุดการสอน i^{th} เวกเตอร์ที่วางอยู่ข้างหนึ่งของเส้นแบ่งแยกฉลากจะเป็น -1 และ เวกเตอร์ที่วางอีกข้างหนึ่งของเส้นแบ่งแยกฉลากจะเป็น 1 ชุดการสอนที่วางอยู่ทั้งสองข้างของเส้นประถูกเรียกว่า เวกเตอร์สนับสนุน (Support vector)

ในการค้นหาเส้นแบ่งแยกที่ดีที่สุดนั้นควรจะเปลี่ยนสมการไปสู่รูปแบบเลอเกรนจ์(Lagrangian formulation) โดยข้อจำกัดบนเลอเกรนจ์มีลติพายเออร์ (Lagrange multipliers) (α_i) จะง่ายกว่าในการแก้สมการรูปแบบเลอเกรนจ์สามารถถูกเขียนให้อยู่ในรูปแบบสมการที่ 2.6

$$L_P \equiv \frac{1}{2} \|w\|^2 - \sum_{i=1}^l \alpha_i y_i (x_i \cdot w + b) + \sum_{i=1}^l \alpha_i \quad (2.6)$$

L_P ควรจะมีค่าน้อยที่สุด ซึ่งคือปัญหาการหาค่าที่ดีที่สุดของปัญหาพหุนามกำลังสอง (Quadratic Programming) ซึ่งใช้เวลานานในการแก้ปัญหา ดังนั้นจะใช้ Sequential Minimal Optimization (SMO) ซึ่งเป็นอัลกอริทึมสำหรับการสอนซัพพอร์ตเวกเตอร์แมชชีนในกรณีที่มีปัญหาพหุนามกำลังสองขนาดใหญ่ถูกแตกออกเป็นชุดปัญหาเล็กที่สุดเท่าที่จะเป็นไปได้ซึ่งสามารถถูกแก้ไขได้ SMOสามารถจัดการกับชุดข้อมูลที่ใช้สอนที่มีขนาดใหญ่ภายใต้เวลาที่เหมาะสม ปัญหาคู่ (Dual problem) สามารถถูกแก้ไขปัญหาแทนที่ของ L_P โดย L_D เป็นผลลัพธ์โดยการแทนที่ L_P ซึ่งแสดงได้ดังสมการที่ 2.7, 2.8 และ 2.9

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \quad (2.7)$$

$$w = \sum_i \alpha_i y_i x_i \quad (2.8)$$

$$\sum_i \alpha_i y_i = 0 \quad (2.9)$$

ซัพพอร์ตเวกเตอร์แมชชีนใช้ประเภทพิเศษของคลาสฟังก์ชัน (Function class) ซึ่งใช้โดยใช้เคอร์เนล (Kernel) เป็นตัวจำแนกเส้นขอบขนาดใหญ่ของพื้นที่คุณลักษณะ เคอร์เนล $k(x, y)$ เป็นการวัดความคล้ายคลึงกันซึ่งกำหนดโดยการจับคู่แบบไม่ชัดแจ้ง (Implicit mapping) ϕ , จากพื้นที่จุดกำเนิดไปสู่พื้นที่คุณลักษณะจนกระทั่ง $k(x, y) = \phi(x) \cdot \phi(y)$ โดยรูปแบบปกติซัพพอร์ตเวกเตอร์แมชชีนเรียนรู้จากฟังก์ชันเส้นตรง (Linear function) ถ้าไม่เช่นนั้นแล้วซัพพอร์ตเวกเตอร์แมชชีนจะการเรียนรู้ด้วยตัวจำแนกแบบโพลิโนเมียลและเรเดียลฟังก์ชัน (Polynomial and radial basis function classifiers) การเลือกฟังก์ชันเคอร์เนล (Kernel function) เป็นสิ่งที่สำคัญมากในการจำแนกอย่างมีประสิทธิภาพของซัพพอร์ตเวกเตอร์แมชชีน ตัวอย่างของเคอร์เนลได้แก่

- เคอร์เนลเส้นตรง (Linear kernel) : $k(x,y) = x^T y$
- เคอร์เนลโพลิโนเมียล (Polynomial kernel): $k(x,y) = (\mathbf{V}(x^T y) + c)d$

- เคอร์เนลเรเดียลฟังก์ชัน (Radial basis function kernel) : $k(x,y) = \exp(-\gamma \|x-y\|^2)$

ในงานวิจัยนี้ผู้วิจัยได้ใช้ SVM^{Light} (Thorsten, 1999) เป็นเครื่องมือสำหรับการจำแนกหัวเรื่องหัวเรื่องจากรูปแบบของการแสดงผลโดยการใช้ซัพพอร์ตเวกเตอร์แมชชีน

2.2 งานวิจัยที่เกี่ยวข้อง

ในส่วนของงานวิจัย และบทความที่เกี่ยวข้องกับวิทยานิพนธ์ฉบับนี้ ประกอบไปด้วย งานวิจัยที่เกี่ยวกับประโยชน์ของหัวเรื่อง, งานวิจัยเกี่ยวกับการแบ่งเว็บเพจเป็นส่วนย่อย, งานวิจัยเกี่ยวกับการจำแนกหัวเรื่อง

2.2.1 งานวิจัยที่เกี่ยวกับประโยชน์ของหัวเรื่อง

Watanabe (2007) ได้เสนอการศึกษาความง่าย (Usability) ในการใช้งานเว็บของผู้ใช้ที่มีสายตาปกติกับผู้พิการทางสายตาในการใช้เว็บไซต์ที่หัวเรื่องใช้ป้ายระบุหัวเรื่องในการกำหนดรูปแบบของการแสดงผลเปรียบเทียบกับเว็บไซต์ที่หัวเรื่องไม่ได้ใช้ป้ายระบุหัวเรื่องในการกำหนดรูปแบบของการแสดงผล ผลการวิจัยพบว่าเว็บไซต์ที่หัวเรื่องมีการใช้ป้ายระบุหัวเรื่องในการกำหนดรูปแบบของการแสดงผลมีความง่ายในการใช้งานมากกว่าเว็บที่หัวเรื่องไม่ได้ใช้ป้ายระบุหัวเรื่องในการกำหนดรูปแบบของการแสดงผล และผู้พิการทางสายตาได้รับประโยชน์มากกว่าผู้ที่มีสายตาปกติในกรณีที่หัวเรื่องของเว็บไซต์ใช้ป้ายระบุหัวเรื่องในการกำหนดรูปแบบของการแสดงผล

2.2.2 งานวิจัยเกี่ยวกับการแบ่งเว็บเพจเป็นส่วนย่อย

Ramakrishnan (2004) ได้เสนอเว็บเบราว์เซอร์ซึ่งแบ่งเว็บเพจด้วยการวิเคราะห์ทางความหมายและโครงสร้าง เพื่อแปลงเอกสารเอชทีเอ็มแอลไปสู่โครงสร้างทางความหมายในรูปแบบของวอยซ์เอกซ์เอ็มแอลทรี (VoiceXML Tree) ซึ่งทำให้ผู้ใช้สามารถเข้าไปสู่เนื้อหาที่ต้องการในแนวทางตามความหมายของเนื้อหา โดยใช้วิธีในการตรวจสอบแบบแผนและการวิเคราะห์ข้อความเพื่อค้นหาส่วนย่อยที่มีความหมายเหมือนกัน

Baluja (2006) ได้เสนอการแบ่งเว็บเพจเป็นส่วนย่อยโดยใช้ต้นไม้ตัดสินใจเป็นเครื่องมือในการแบ่งเว็บเพจออกเป็น 9 ส่วนที่เหมาะสมสำหรับการใช้เว็บเพจผ่านอุปกรณ์ที่มีหน้าจอขนาดเล็ก โดยแสดงผลเป็นเว็บเพจเป็นรูปภาพขนาดเล็ก และผู้ใช้กดแป้นพิมพ์ด้วยตัวเลข 1-9 ของอุปกรณ์ที่มีหน้าจอขนาดเล็กเพื่อเลือกพื้นที่ต้องการขยายเพื่อดูรายละเอียดเพิ่มเติมเกี่ยวกับส่วนย่อยนั้น

Jiuxin (2007) ได้เสนอการแบ่งเว็บเพจเป็นส่วนย่อยโดยใช้การประมวลผลภาพ (Image processing) และคุณลักษณะพิเศษของเว็บเพจ เว็บเพจจะถูกแปลงไปเป็นรูปภาพและทำการย่อและแบ่งเป็นรอบๆ ตามเงื่อนไขที่กำหนด ซึ่งเว็บเพจที่ถูกแปลงเป็นรูปภาพจะถูกแบ่งเป็นรูปภาพที่มีขนาดเล็กจนได้ส่วนย่อยของเว็บเพจที่ต้องการ

Guo (2007) ได้เสนอการแบ่งเว็บเพจเป็นส่วนย่อยโดยใช้ข้อมูลตำแหน่งการแสดงผล (Visual layout information) ซึ่งได้จากมอซิลลาเฟรมทรี (Mozilla frame tree) แทนที่จะใช้ข้อมูลที่ได้จากคอมในการพิจารณาตำแหน่งที่ตั้งที่เกี่ยวกับระยะ (Spatial locality) โดยใช้ตัวแบ่งแยกซึ่งถูกระบุด้วยป้ายระบุเขตที่เอมแอลและข้อมูลตำแหน่งการแสดงผลมาทำการแบ่งแยกเว็บเพจเป็นส่วนย่อย

2.2.3 งานวิจัยที่เกี่ยวกับการจำแนกหัวเรื่อง

Watanabe et al. (2007) ได้เสนอวิธีในการจำแนกหัวเรื่องที่ไม่ได้ใช้ป้ายระบุหัวเรื่องในการกำหนดรูปแบบของการแสดงผลโดยใช้คนเป็นผู้ทดลอง เพื่อศึกษาว่าคนจะใช้ลักษณะเด่นใดบ้างในการระบุว่าส่วนใดของข้อความที่ปรากฏในหน้าเว็บเพจเป็นหัวเรื่อง พบว่าผู้ทดลองสามารถแยกแยะข้อความที่เป็นหัวเรื่องออกจากเว็บเพจได้จากลักษณะเด่นดังต่อไปนี้ ขนาดของตัวอักษร, สีของตัวอักษร, สีพื้นหลังตัวอักษร, การใช้ไอคอน (Icon) หน้าข้อความที่เป็นหัวเรื่อง และการไม่ย่อหน้าของข้อความ แต่งานวิจัยนี้ยังต้องใช้คนในการจำแนกหัวเรื่องที่ไม่ได้ใช้ป้ายระบุหัวเรื่องในการกำหนดรูปแบบของการแสดงผล

Brudvik et al. (2008) ได้เสนอระบบ HeadingHunter ซึ่งเป็นวิธีการจำแนกหัวเรื่องที่ได้จากรูปแบบของการแสดงผลโดยใช้ต้นไม้ตัดสินใจ (Decision tree) และใช้ลักษณะเด่นดังนี้ ขนาดตัวอักษร, ความหนาของตัวอักษร, ความยาวของข้อความ, อัตราส่วนความกว้างเทียบกับความสูง แต่งานวิจัยนี้ไม่ได้ใช้ลักษณะเด่นของจำแนกหัวเรื่องที่ได้จากการศึกษาว่าคนใช้ลักษณะเด่นอะไรในการระบุว่าข้อความใดเป็นหัวเรื่อง ทำให้การจำแนกหัวเรื่องของงานวิจัยนี้ยังไม่ครอบคลุมการ

จำแนกข้อความที่เป็นหัวเรื่องทั้งหมด และไม่ได้กำหนดขอบเขตของการหาหัวเรื่องของส่วนย่อยที่ชัดเจนทำให้มีหัวเรื่องของส่วนย่อยอื่นรวมเข้ามาในขอบเขตของการหาหัวเรื่องของส่วนย่อยนั้นทำให้การจำแนกหัวเรื่องไม่มีความแม่นยำเท่าที่ควร

ตารางที่ 2.1

แสดงการเปรียบเทียบข้อดี-ข้อเสียของเทคนิคต่างๆ ที่ใช้ในจำแนกหัวเรื่อง

งานวิจัย	เทคนิคที่ใช้	ข้อดี	ข้อเสีย
Watanabe et al. (2007)	ใช้คนในการระบุ ลักษณะเด่นของใน การจำแนกหัวเรื่อง	ลักษณะเด่น ครอบคลุมการระบุหัว เรื่อง	ไม่สามารถจำแนกหัว เรื่องแบบอัตโนมัติได้
Brudvik et al. (2008)	ใช้ต้นไม้ตัดสินใจเพื่อ สร้างแบบจำลองใน การจำแนกหัวเรื่อง	จำแนกหัวเรื่องแบบ อัตโนมัติ	ลักษณะเด่นไม่ ครอบคลุมการจำแนก หัวเรื่องทั้งหมดและไม่ มีการกำหนดขอบเขต การหาหัวเรื่องที่ แน่นอน