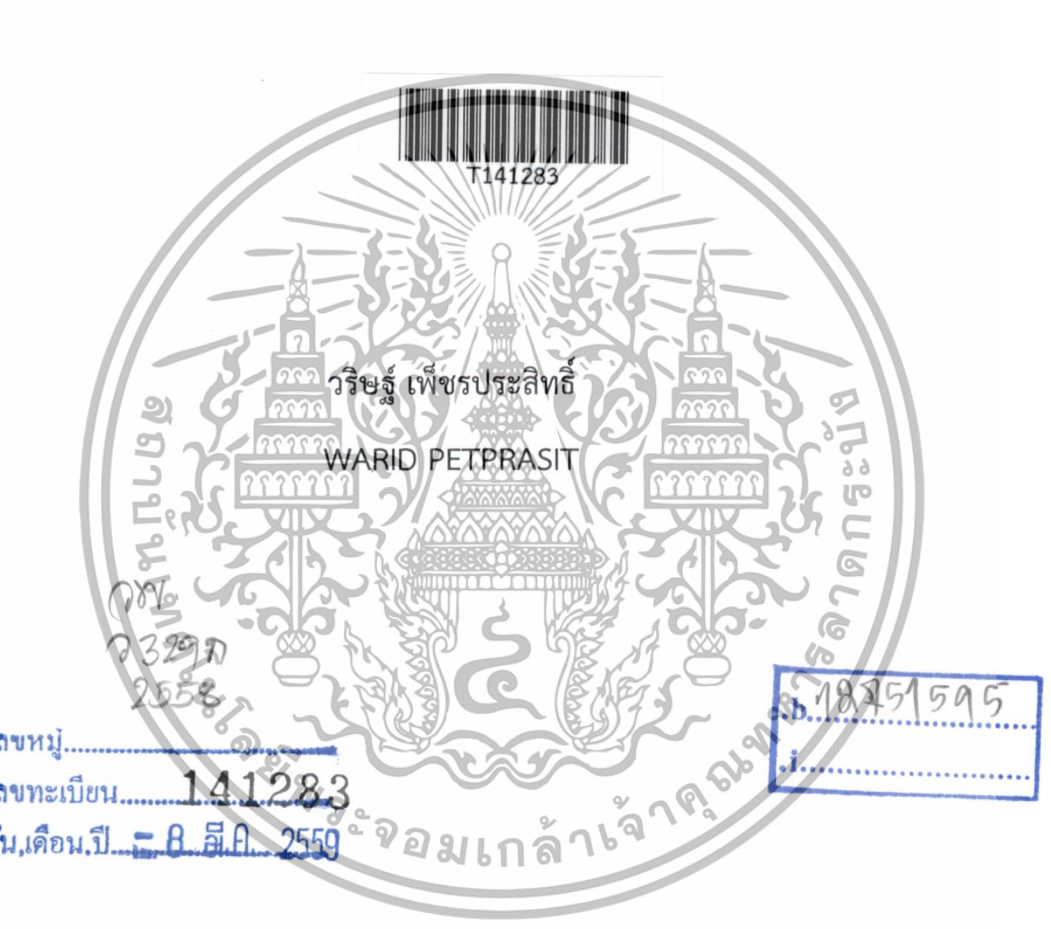
การสกัดเนื้อหาเว็บโดยใช้การตรวจจับหัวข้อ และความหนาแน่นของโหนด

WEB CONTENT EXTRACTION BASED ON SUBJECT DETECTION AND NODE DENSITY



วริษฐ์ เพ็ชรประสิทธิ์

WARID PETPRASIT

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2558

KMITL-2015-SC-M-002-049

WEB CONTENT EXTRACTION BASED ON SUBJECT DETECTION AND NODE DENSITY

WARID PETPRASIT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2014

KMITL-2015-SC-M-002-049

คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ "การสกัดเนื้อหาเว็บโดยใช้การตรวจจับหัวข้อและความหนาแน่นของโหนด "
"WEB CONTENT EXTRACTION BASED ON SUBJECT DETECTION AND NODE DENSITY"

ชื่อนักศึกษา นายวริษฐ์ เพ็ชรประสิทธิ์

รหัสประจำตัว 566050804

ปริญญา วิทยาศาสตรมหาบัณฑิต (สาขาวิชาวิทยาการคอมพิวเตอร์)

ภาควิชา วิทยาการคอมพิวเตอร์

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ดร.สายชล ใจเย็น

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม -

คณะกรรมการสอบวิทยานิพนธ์	ลายมือชื่อ
ผศ.ดร.ศรัณย์ อินทโกสุม ประธานกรรมการ	
ผศ.ดร.อนันตพร หรรษคุณาชัย อาจารย์บัณฑิตประจำ (ในสาขาวิชาที่เกี่ยวข้อง)	
ดร.ธนัสนี เพียรตระกูล ผู้ทรงคุณวุฒิจากภายนอกสถาบันฯ	
ดร.สายชล ใจเย็น อาจารย์ที่ปรึกษาวิทยานิพนธ์	

วัน/ เดือน/ ปี ที่สอบ วันที่ 30 มิถุนายน พ.ศ.2558

สถานที่สอบ ห้อง 306 อาคารปฏิบัติการใหม่ ชั้น 3

คณะวิทยาศาสตร์รับรองแล้ว

(รองศาสตราจารย์ ดร.ดุษณี ธนะบริพัฒน์)
คณบดีคณะวิทยาศาสตร์
วันที่ 18 เดือน ก.ค. พ.ศ. 58

หัวข้อวิทยานิพนธ์	การสกัดเนื้อหาเว็บโดยใช้การตรวจจับหัวข้อและความหนาแน่นของโหนด WEB CONTENT EXTRACTION BASED ON SUBJECT DETECTION AND NODE DENSITY
นักศึกษา	นายวริษฐ์ เพ็ชรประสิทธิ์
รหัสประจำตัว	55650804
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ภาควิชา	วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์	อาจารย์ ดร.สายชล ใจเย็น

บทคัดย่อ

ในปัจจุบัน มีข้อมูลปริมาณมหาศาลถูกส่งมาจากทุกหนทุกแห่งผ่านทางโลกออนไลน์ ด้วยเหตุนี้นักวิจัยหลายกลุ่มจึงหันมาสนใจการสกัดข้อมูลจากหน้าเว็บ โดยนักวิจัยส่วนใหญ่ได้มุ่งไปที่การคัดกรองเนื้อหาที่เป็นประโยชน์จากหน้าเว็บไซต์ ปัญหาหลักของงานด้านนี้คือความยืดหยุ่นของโมเดล โดยโมเดลส่วนใหญ่นั้นจำเป็นต้องประกาศกฎการสกัดข้อมูลใหม่เมื่อต้องการสกัดข้อมูลจากหน้าเว็บของเว็บไซต์โดเมนอื่นที่มีโครงสร้างต่างออกไป ถึงแม้ว่า จะมีนักวิจัยบางกลุ่มเสนอการสกัดข้อมูลแบบอัตโนมัติขึ้นมาเพื่อเพิ่มความยืดหยุ่นของโมเดลเหล่านี้ แต่ก็ยังมีข้อจำกัดตรงที่มันสามารถทำงานได้เฉพาะบนหน้าเว็บนำเสนอลิ้งค์เท่านั้น (Link Offer Page) จึงไม่สอดคล้องกับวัตถุประสงค์ของงานบางประเภท เช่น การเก็บรายละเอียดผลิตภัณฑ์มือสอง ต่อมา ได้มีการพัฒนาระบบการสกัดข้อมูลโดยใช้ความหนาแน่นของตัวอักษรในการระบุโหนดข้อมูล ถึงแม้ว่าจะสามารถทำงานแบบอัตโนมัติได้ แต่ก็ยังมีปัญหากับเว็บไซต์ประเภทซื้อขายสินค้าออนไลน์อยู่ดี เนื่องจากเว็บไซต์ประเภทนี้จะมีความหนาแน่นของตัวอักษรในแต่ละโหนดต่ำ และในโหนดเนื้อหาก็ยังมีโหนดอื่นอยู่ปริมาณมาก เช่น โหนดราคาผลิตภัณฑ์ เป็นต้น ดังนั้น งานวิจัยนี้จึงได้นำเสนอวิธีการสกัดข้อมูลจากหน้าเว็บโดยใช้การตรวจจับโหนดหัวข้อและความหนาแน่นของโหนดเพื่อแก้ปัญหาดังกล่าว

คำสำคัญ : การจัดเตรียมข้อมูล การสกัดข้อมูลจากเว็บไซต์ ความหนาแน่นของโหนด ตรวจหาโหนดหัวข้อ พาณิชย์อิเล็กทรอนิกส์ หน้าเว็บที่มีข้อมูลอยู่อย่างหนาแน่น

Thesis Title	WEB CONTENT EXTRACTION BASED ON SUBJECT DETECTION AND NODE DENSITY
Student	Warid Petprasit
Student ID	55650804
Degree	Master of Science
Program	Computer Science
Year	2015
Thesis Advisor	Dr. Saichon Jaiyen

Abstract

Currently, a large amount of data has been transferred from everywhere through World Wide Web. For this reason, many researchers have interested in web information extraction. They focused on filtering only useful data from web pages. Generally, a major problem of the web extraction software is model flexibility, since it was necessary to redefine rules of software when it operated on cross domains with different structure. Even though some researchers tried to improve the flexibility by presenting the web extraction in automatic manner, the model was limited to operating only on link offer page. So, it is inconsistent with objectives of some websites like second hand sales. Subsequently, some of them developed web extraction systems by identifying content node with text density. The systems can automatically extract the web information; however, they are incompatible with E-Commerce web pages. This is because the pages have low text density per node and also contain many nodes, i.e. price, in content node. Consequently, we propose web content extraction based on subject detection and node density for solving the mentioned problems.

Keywords : data preparation web information extraction node density subject detection e-commerce data intensive

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จลุล่วงไปด้วยดี เนื่องจากผู้จัดทำได้รับความช่วยเหลือจากบุคคลผู้มีพระคุณหลายท่าน ดังนี้

ขอขอบพระคุณ ดร.สายชล ใจเย็น อาจารย์ประจำสาขาวิชาวิทยาการคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้สละเวลามาช่วยตรวจทานแก้ไข และเพิ่มเติมในทุกส่วนของเอกสารต่างๆ พร้อมทั้งได้ช่วยเพิ่มเติมแนวคิดบางส่วนให้โมเดลนี้มีประสิทธิภาพมากขึ้น

ขอขอบพระคุณ ผศ.ดร.นวลสวาท หิรัญสกลวงศ์ อาจารย์ประจำสาขาวิชาวิทยาการคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ผู้ให้คำปรึกษาในช่วงต้นของการทำงาน และเสริมแนวคิดในเรื่องเกี่ยวกับ การค้นคืนข้อมูล (Information Retrieval)

ขอขอบพระคุณ อาจารย์ผู้สอนรายวิชาต่างๆให้กับข้าพเจ้า ทำให้ระหว่างศึกษาในรายวิชาต่างๆนั้น ได้สามารถนำมาปรับและประยุกต์ใช้ในงานวิจัยชิ้นนี้ได้ โดยเฉพาะ ผศ.ดร.อนันตพร หรรษคุณาฒัย ในรายวิชา Data Mining Theory

ขอขอบพระคุณ อาจารย์ผู้ให้คำแนะนำ และคำปรึกษาในรายวิชาสัมมนาทุกท่านที่สละเวลามาช่วยแสดงความคิดเห็นและชี้จุดเด่น จุดด้อยของหัวข้อและงานวิจัยนี้ โดยมี ผศ.ดร.ศรัณย์ อินทโกสุม รศ.ดร.วีระ บุญจริง ดร.สุวรรณ จันทิวาสารกิจ ผศ.ดร.นันทิกา เบญจเทพานันท์ ผศ.สิริลักษณ์ อนันต์สถิตย์สิน และ รศ.ธีรวัฒน์ ประกอบผล

ท้ายนี้ ขอขอบพระคุณบิดา มารดา และญาติพี่น้องที่ได้ให้กำลังใจ และทุนการศึกษาเพื่อให้ศึกษาและสำเร็จลุล่วงไปด้วยดี

นายวริษฐ์ เพ็ชรประสิทธิ์

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันอินเตอร์เน็ตได้รับความนิยมอย่างมาก จึงทำให้การศึกษาค้นคว้าของนักวิจัยส่วนมากมุ่งไปที่อินเตอร์เน็ต ซึ่งองค์ประกอบที่สำคัญของอินเตอร์เน็ตนั้นก็คือเว็บไซต์โดยเป็นเอกสารสำหรับเผยแพร่ในที่สาธารณะในหลากหลายวัตถุประสงค์ ตามแต่ผู้ที่จัดทำขึ้นต้องการจะสื่อสาร ด้วยเหตุผลเหล่านี้จึงเกิดความต้องการในการสร้างเว็บไซต์ขึ้นมาอย่างมากมาย เนื่องจากสามารถเผยแพร่ออกสู่สาธารณะได้ง่าย จึงเกิดปัญหาตามมาเกี่ยวกับการที่มีข้อมูลอยู่ปริมาณมากบนเว็บไซต์ต่างๆ ปะปนอยู่กับโฆษณาหรือสิ่งอื่นๆ ที่ไม่ใช่เนื้อหารายละเอียดที่ผู้เยี่ยมชมต้องการ จึงได้เกิดงานวิจัยทางด้าน การสกัดข้อมูลจากหน้าเว็บ (Web Information Extraction) ขึ้นหลากหลายชิ้นงาน

ระบบการสกัดข้อมูลจากเว็บไซต์ เป็นประโยชน์ในการจัดเตรียมข้อมูล (Pre-Processing) สำหรับงานวิจัยทางด้านการทำเหมืองข้อมูลเว็บไซต์ (Web Mining) ยกตัวอย่างเช่น ระบบการแนะนำผลิตภัณฑ์หรือบริการ ระบบสนุบสนุนการตัดสินใจ ระบบผู้เชี่ยวชาญ ระบบการค้นหาองค์ความรู้ และอีกหลายๆเรื่องที่เป็นงานเฉพาะทาง เช่น การตรวจจับการโฆษณาหรือสิ่งผิดกฎหมายบนเว็บไซต์ และการพยากรณ์ความต้องการ เป็นต้น ดังนั้นจึงมีงานวิจัยหลายงานที่มุ่งเน้นมาที่งานประเภทนี้ เนื่องจากโครงสร้างของหน้าเว็บไซต์ขึ้นกับนักพัฒนาเว็บไซต์และภาษาโปรแกรมที่ใช้ จึงมีการจัดวางโครงสร้างองค์ประกอบของหน้าเว็บที่แตกต่างกัน จึงเป็นการยากที่โมเดลจะสามารถรองรับหน้าเว็บไซต์ในรูปแบบต่างๆได้ และเป็นการยากที่จะสามารถสกัดข้อมูลออกมาแบบอัตโนมัติ เพื่อลดภาระของผู้ใช้งาน นอกจากนี้ยังมีงานวิจัยหลายงานมุ่งไปที่การสกัดเนื้อหาจากหน้ารวมลิ้งค์จากเว็บไซต์ประเภทซื้อขายสินค้า (E-Commerce) และมีเพียงบางงานวิจัยที่มุ่งเน้นไปที่หน้าเว็บที่มีรายละเอียดอยู่อย่างหนาแน่น (Data Intensive) ซึ่งในงานวิจัยเหล่านี้ก็ยังติดปัญหาที่ความเป็นอัตโนมัติเพราะต้องการให้ผู้ใช้งานระบุหัวข้อก่อนการสกัดเนื้อหา หรือ ระบุคำสำคัญที่ใช้ในการสกัดเนื้อหา แล้วโมเดลจึงนำไปแปลงเป็น นิพจน์ปรกติ (Regular Expression) เพื่อทำการสกัดข้อมูลอีกทีหนึ่ง ดังนั้นในงานวิจัยนี้จึงมุ่งไปที่การสกัดข้อมูลจากหน้าเว็บประเภท ซื้อ-ขาย สินค้า ในหน้ารายละเอียดผลิตภัณฑ์ โดยใช้วิธีการตรวจจับโหนดหัวข้อ และความหนาแน่นของโหนดเพื่อระบุโหนดเนื้อหาแบบอัตโนมัติ

1.2 วัตถุประสงค์ของงานวิจัย

1) สกัดโหนดเนื้อหาจากหน้าเว็บไซต์ซื้อขายสินค้าออนไลน์ ในหน้าที่มีรายละเอียดอยู่อย่างหนาแน่นแบบอัตโนมัติ
2) พัฒนาอัลกอริธึมที่ใช้สำหรับสกัดโหนดเนื้อหาของเว็บไซต์ประเภทซื้อขายสินค้าออนไลน์ ในหน้าเว็บที่มีข้อมูลอยู่อย่างหนาแน่น (Data Intensive Page) แบบอัตโนมัติ

1.3 ขอบเขตของงานวิจัย

1) สกัดส่วนของเนื้อหาจากหน้าเว็บไซต์ประเภทซื้อขายสินค้าออนไลน์
2) สกัดในส่วนของหน้าที่มีข้อมูลอยู่หนาแน่น
3) สกัดข้อมูลจากหน้าเว็บที่มีโหนดหัวข้อบนโครงสร้างเอชทีเอ็มแอล

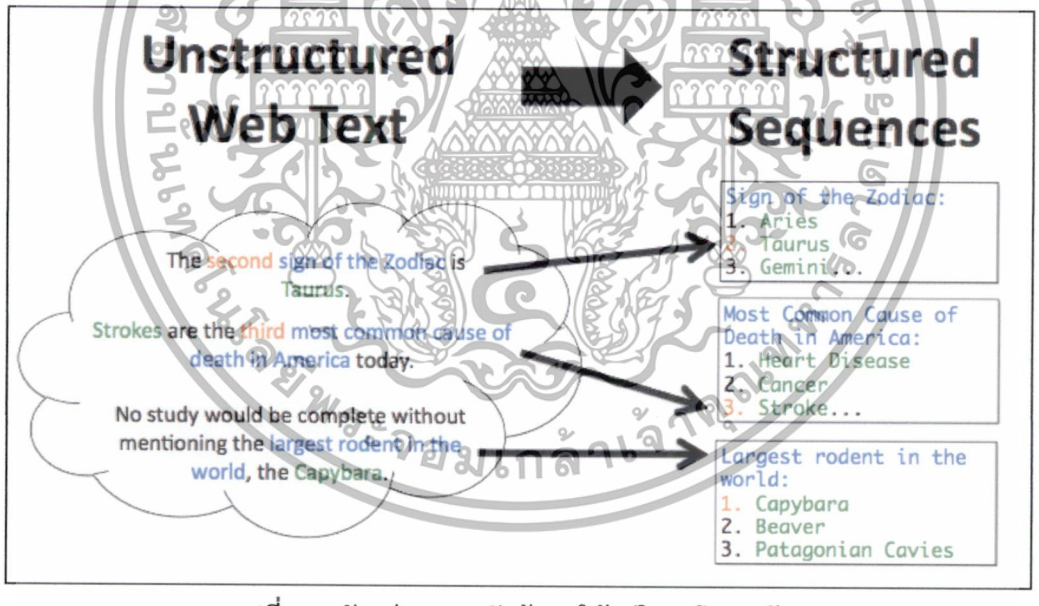
1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) สามารถรองรับการทำงานกับเว็บไซต์ได้หลากหลายรูปแบบ
2) สามารถรองรับการทำงานกับเว็บไซต์ ซื้อ-ขาย สินค้าที่ได้รับความนิยมสูงได้
3) สามารถดัดแปลงหรือปรับแต่งค่าพารามิเตอร์เพื่อให้เหมาะสมกับการนำไปใช้งานได้ง่าย
4) สามารถลดภาระในการเก็บข้อมูลจากเว็บไซต์เพื่อนำไปวิเคราะห์ต่อได้

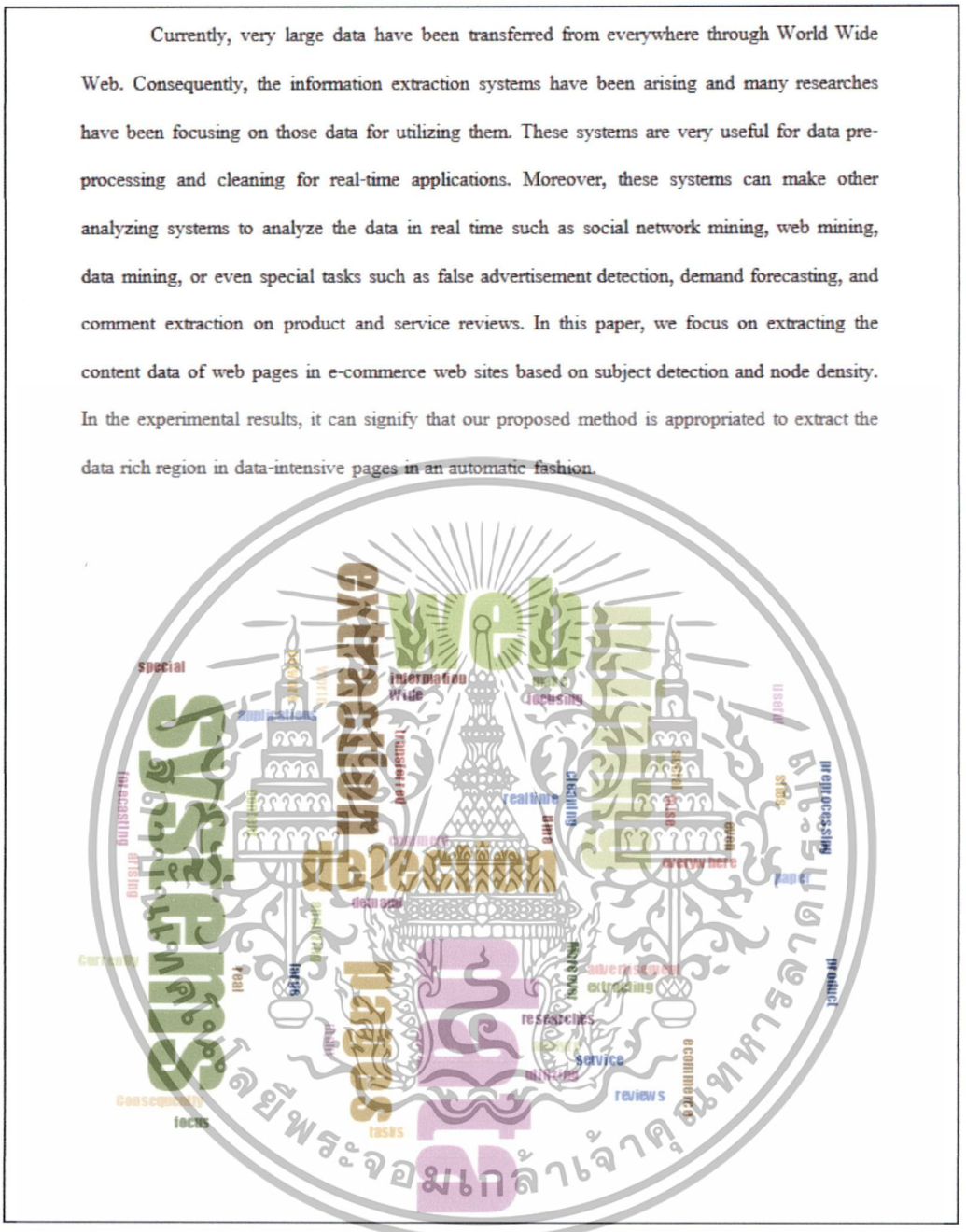
บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การสกัดข้อมูล (Information Extraction)

การสกัดข้อมูล คือการดึงข้อมูลเฉพาะส่วนที่ต้องการออกมาในรูปแบบที่ถูกจัดโครงสร้างแล้ว เช่น XML หรือ JSON เป็นต้น โดยการสกัดข้อมูลส่วนใหญ่จะทำงานกับข้อมูลที่ยังไม่ถูกจัดโครงสร้าง หรือข้อมูลที่อยู่ในรูปแบบกึ่งโครงสร้าง ดังรูปที่ 2.1 ซึ่งเป้าหมายของงานประเภทนี้คือการทำให้กระบวนการสกัดข้อมูลเป็นแบบอัตโนมัติมากที่สุด เพื่อที่จะนำไปใช้ประโยชน์ในงานด้านอื่นๆ ต่อไป ยกตัวอย่างเช่น รูปที่ 2.2 เป็นตัวอย่างของการสร้างป้ายคำสำคัญแบบก้อนเมฆที่ถูกสร้างจากบทคัดย่อภาษาอังกฤษในงานวิจัยนี้ผ่านเว็บไซต์ (https://www.jasondavies.com/wordcloud) โดยวิธีการนับความถี่ของคำศัพท์ที่ใช้ และทำการเพิ่มขนาดของคำ หลังจากนั้นจึงทำการตกแต่งด้วยสีและการเปลี่ยนองศาของคำ ซึ่งสามารถพบได้บ่อยในเว็บไซต์ประเภทบทความทั่วไป ที่ใช้ป้ายกำกับคำในการระบุประเภทของบทความ

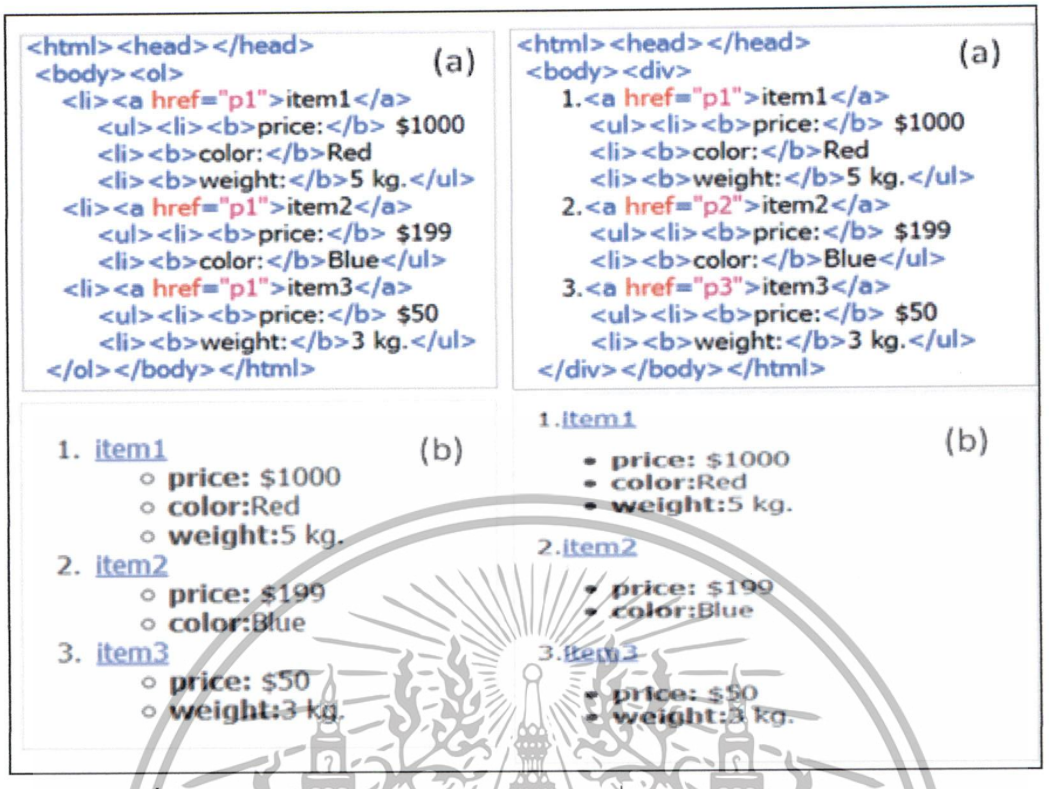


รูปที่ 2.1 ตัวอย่างการสกัดข้อมูลให้อยู่ในรูปโครงสร้าง



รูปที่ 2.2 ตัวอย่างป้ายคำสำคัญแบบก้อนเมฆ (Tag Cloud)

การสกัดข้อมูลที่ได้รับความนิยมอย่างมากในปัจจุบัน คือการสกัดข้อมูลจากหน้าเว็บเพจ ซึ่งปัจจุบันมีบริการประเภทนี้เยอะขึ้น และทำให้การเก็บดึงข้อมูลจากหน้าเว็บเป็นเรื่องที่ง่ายขึ้น โดยการสกัดข้อมูลนั้นเป็นกระบวนการพื้นฐานก่อนการทำงานด้านอื่นๆ เช่น การจัดทำดัชนีจากเอกสาร (Indexing) การสรุปใจความสำคัญ (Text Summarization) ระบบผู้เชี่ยวชาญ (Expert System) การแปลงข้อมูลจากรูปแบบหนึ่งไปเป็นอีกรูปแบบหนึ่ง (Transformation) เพื่อความง่ายต่อการนำไปประยุกต์ใช้งาน



รูปที่ 2.3 ตัวอย่างของ HTML Source Code ที่ต่างกัน แต่แสดงผลเหมือนกัน

ในงานวิจัยนี้ได้มุ่งเน้นไปที่การสกัดข้อมูลจากหน้าเว็บ (Web Information Extraction) เพราะบนหน้าเว็บแต่ละหน้าเว็บนั้น มีโครงสร้างที่แตกต่างกัน แต่ผลลัพธ์ในการแสดงผลบนหน้าเว็บนั้นอาจจะออกมาเหมือนกันหรือคล้ายกันดังรูปที่ 2.3 ซึ่งเป็นผลมาจากความแตกต่างด้านประสบการณ์ของนักพัฒนาเว็บไซต์ ภาษาโปรแกรมที่ใช้พัฒนาเว็บไซต์ ลักษณะความชอบเฉพาะตัวของนักพัฒนาเว็บไซต์ วัตถุประสงค์ของการพัฒนาเว็บไซต์ รวมถึงวัตถุประสงค์ของผู้ใช้งานหรือผู้บริโภคนั่นเอง ยกตัวอย่างเช่น เว็บไซต์ที่มีวัตถุประสงค์เพื่อขายสินค้า ก็จะพยายามจัดเรียงสินค้าให้ดูง่ายและเหมาะสมกับการเรียกดูของผู้ใช้งาน มองเห็นได้หลายชิ้นในหน้าเดียว ซึ่งในสายตาของผู้ใช้งานนั้นอาจดูคล้ายกันมาก แต่เมื่อดูเบื้องหลังของการพัฒนาเว็บไซต์หรือส่วนที่เรียกว่า ซอร์สโค้ด (Source Code) หรือรหัสต้นฉบับของเว็บเพจนั้นๆ จะมีความต่างกันโดยสิ้นเชิง

ในงานวิจัยนี้จะยกตัวอย่างของเทคนิคที่นิยมใช้ในการสกัดข้อมูลซึ่งมีอยู่ 3 วิธีการหลักด้วยกัน คือ วิธีที่ขึ้นกับการประมวลภาษาธรรมชาติ (Natural Language Processing) วิธีที่ขึ้นกับแบบจำลองเอกสารเชิงวัตถุ (Document Object Model) และวิธีที่ขึ้นกับการมองเห็นของมนุษย์ (Vision Based) ดังนี้

2.1.1 วิธีที่ขึ้นกับการประมวลผลภาษาธรรมชาติ (NLP Based)

การประมวลผลภาษาทางธรรมชาติ (NLP) เป็นสาขาหนึ่งของปัญญาประดิษฐ์ โดยจะใช้งานในเรื่องของการทำให้เครื่องคอมพิวเตอร์สามารถเข้าใจภาษาของมนุษย์ได้ โดยงานวิจัยทางด้านนี้หากมีการทำงานบนหน้าเว็บ จะถูกเรียกอีกชื่อหนึ่งว่า Web Semantic ยกตัวอย่างเช่น การแบ่งคำ (Word Segmentation) เพื่อนำไปตัดคำหรือประโยคในภาษาต่างๆ โดยภาษาอังกฤษจะใช้การเว้นวรรค แต่ภาษาไทยมีความยากกว่าเนื่องจากการเขียนข้อความเป็นประโยค ส่วนใหญ่จะใช้ความจำความเข้าใจของตัวบุคคล เพื่อทำการแปลความหมายออกมาในรูปของประโยค เช่น คำว่า ตากลม จะสามารถแบ่งคำได้สองแบบคือ ตา-กลม หรือ ตาก-ลม เป็นต้น การตีความหมายของความคิดเห็นว่าเป็นความคิดเห็นในเชิงบวก หรือเชิงลบ (Opinion Mining) ในการติชมจากหน้าเว็บไซต์ต่างๆ [รูปที่ 2.5] ไม่ว่าจะเป็น สถานที่ท่องเที่ยว ที่พัก หรือแม้กระทั่งผลิตภัณฑ์ต่างๆ อีกตัวอย่างหนึ่งที่นิยมนำ NLP ไปใช้แก้ปัญหาคือ การรู้จำอักขระด้วยแสง (Optical Character Recognition) คือ การแปลงรูปภาพของข้อความไปเป็นข้อความเพื่อนำไปใช้งานต่อดังรูปที่ 2.4 เป็นต้น

รูปที่ 2.4 ตัวอย่างผลลัพธ์ที่ได้จากการทำ OCR (Optical Character Recognition)

รูปที่ 2.5 ความคิดเห็นต่อโรงแรมของผู้ใช้งานในเว็บไซต์ Agoda

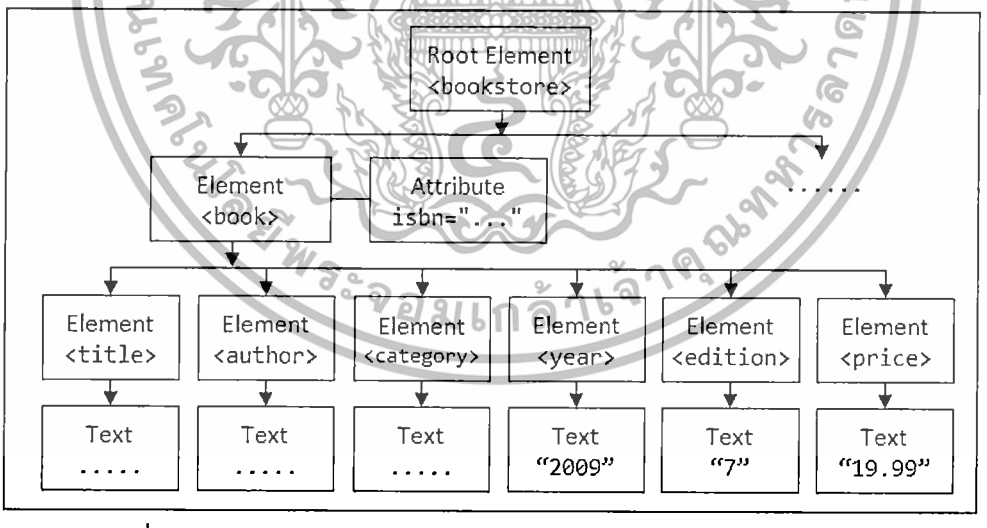
สำหรับในส่วนของงานวิจัยทางด้านการสกัดข้อมูลจากเว็บเพจนี้ ได้มีงานวิจัยหลายงานพยายามนำ NLP ไปใช้ในการตรวจจับข้อความบนหน้าเว็บเพื่อระบุถึงเนื้อหา หรือประเภทของเนื้อหาที่ได้พบโดยใช้ นิพจน์ปรกติ (Regular Expression) ยกตัวอย่างเช่น งานวิจัยของ Zheng, X. Gu, Y. และ Li, Y. [18] ซึ่งกล่าวถึงการสกัดข้อมูลจากหน้าเว็บ โดยขึ้นกับความไร้ระเบียบของข้อมูลบนหน้าเว็บ โดยที่ตัวแปรสำคัญของโมเดลนี้ขึ้นอยู่กับการเลือกพิจารณาคำสำคัญในการระบุประเภทของข้อความก่อน ยกตัวอย่างเช่น นิพจน์ปรกติของโหนดราคา จะต้องขึ้นต้นด้วยตัวเลขและตามด้วยตัวเลขกี่ตัวก็ได้ โดยจะต้องมีจุดทศนิยม 2 ตำแหน่ง และตามด้วยตัวอักษร US ระบบจึงจะระบุว่าโหนดนั้นเป็นโหนดของราคา โดยนิพจน์ปรกติของโหนดราคาที่กล่าวไปแล้วจะแสดงออกมาดังนี้

^\d+(\.\d{2})?$

ข้อดีของการนำนิพจน์ปรกติมาใช้งานนั้น อยู่ที่การไม่ยึดติดในส่วนของโครงสร้างของ HTML ซึ่งนั่นหมายความว่าการสกัดข้อมูลด้วยวิธีนี้ไม่ขึ้นกับโครงสร้างของการพัฒนาเว็บไซต์ เนื่องจากนิพจน์ปรกติจะสามารถดักจับข้อมูลเฉพาะส่วนได้โดยไม่สนใจว่าข้อมูลเบื้องหลัง หรือ รหัสต้นฉบับ (Source Code) จะเป็นอย่างไร และความเร็วของการทำงานถือว่าเร็ว เมื่อเปรียบเทียบกับการทำงานในรูปแบบอื่นๆ และข้อดีที่สามารถเห็นได้ชัดเจนอีกหนึ่งอย่างก็คือ นิพจน์ปรกติสามารถระบุได้ว่า ส่วนของข้อมูลที่ถูกสกัดนั้นมีความหมายว่าอะไร หรือบ่งบอกถึงอะไร และด้วยความที่นิพจน์ปรกตินั้นใช้วิธีการตั้งกฎก่อนที่จะทำการสกัดข้อมูลจึงเกิดปัญหาที่แก้ไขได้ยาก ซึ่งก็คือนิพจน์ปรกติไม่สามารถรู้ได้ว่าส่วนใดเป็นส่วนของรายละเอียดผลิตภัณฑ์ หรือส่วนที่มีต้องการมีลักษณะอยู่ในรูปแบบของบทความหรือย่อหน้าเนื่องจากเป็นส่วนที่ไม่ตายตัว

2.1.2 วิธีที่ขึ้นกับแบบจำลองเอกสารเชิงวัตถุ (DOM Based)

แบบจำลองเอกสารเชิงวัตถุ หรือ Document Object Model (DOM) สำหรับทางด้านเว็บไซต์ คือ โมเดลที่ใช้เพื่อช่วยให้อ่านเอกสาร HTML และ XML โดยแสดงเอกสารจากรูปแบบโครงสร้างออกมาในรูปแบบของโครงสร้างต้นไม้ (Tree Structure) ดังรูปที่ 2.6 ซึ่งยังรวมไปถึงการใช้ DOM เพื่อเดินทางผ่านโหนดต่างๆ ของ แท็ก (Tag) เพื่อกระทำการใดๆ กับโหนดที่ต้องการได้อีกด้วย ยกตัวอย่างเช่น การใช้ภาษาโปรแกรม Java Script และ CSS เพื่อควบคุมการแสดงผลต่างๆ บนหน้าเว็บ โดยผ่านรูปแบบของโครงสร้างต้นไม้ (ในที่นี้จะให้ความหมายของ DOM Tree และ DOM Node เหมือนกัน)



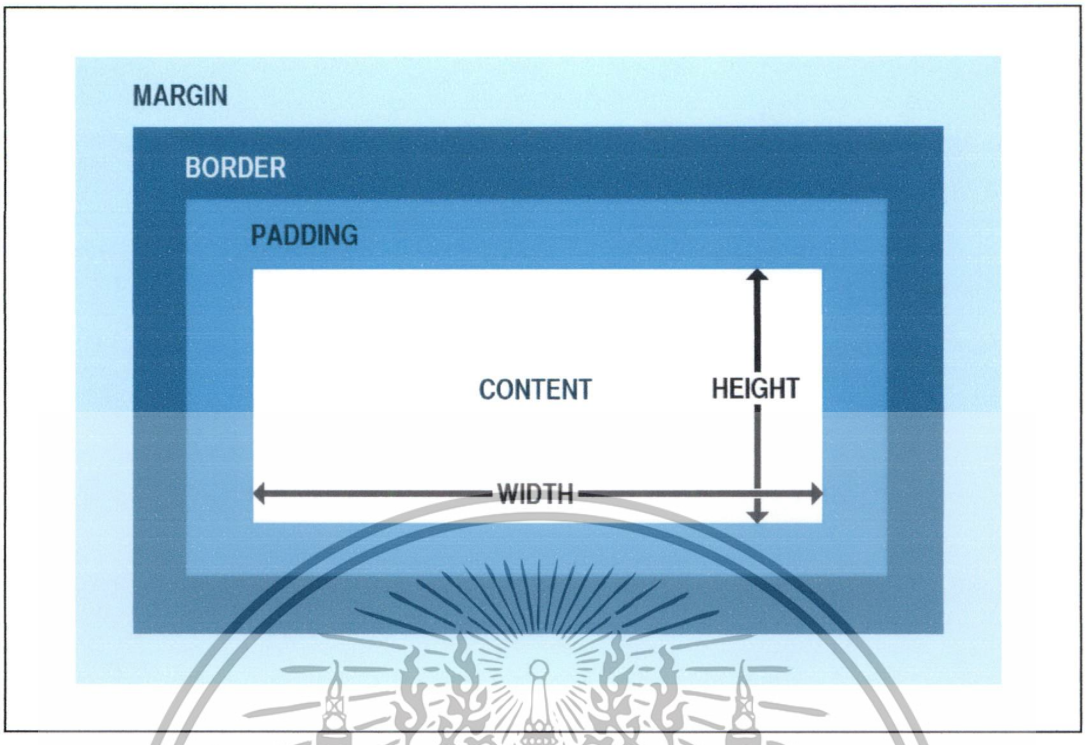
รูปที่ 2.6 โครงสร้างเอกสารเชิงวัตถุ (DOM Node) ของหน้าเว็บแสดงหนังสือ

DOM นั้นได้ถูกนำมาใช้กับงานวิจัยทางด้านการสกัดข้อมูลอย่างแพร่หลาย เนื่องจากเป็นวิธีการพื้นฐานที่ผู้ที่สนใจเกี่ยวกับการพัฒนาเว็บไซต์จำเป็นต้องรู้ และยังสามารถทำความเข้าใจได้ง่าย ดังนั้น DOM จึงได้รับความนิยมสูง โดยที่เกือบจะทุกงานจะมีการคำนึงถึงส่วนประกอบของ DOM เสมอไม่ว่าทางตรงหรือทางอ้อม ซึ่งโดยส่วนใหญ่แล้วการใช้งานโดยมากจะเริ่มที่การแปลงจากโครงสร้าง HTML ให้กลายเป็น DOM ก่อนเริ่มการทำงาน เนื่องจากจะทำให้สามารถเดินทางไปยังโหนดต่างๆ ได้โดยง่ายโดยผ่านโครงสร้างของต้นไม้ดังภาพที่ได้แสดงไปก่อนหน้านี้ และใช้ชื่อและลำดับชั้นของแท็กเป็นดัชนีเพื่อช่วยระบุตำแหน่ง โดยตรงส่วนนี้มีหลายคนให้ความสนใจจึงมีหลายกลุ่มที่สร้างเครื่องมือไว้ให้เรียกใช้งานอยู่ในหลายภาษาโปรแกรม

สำหรับความได้เปรียบของ DOM Model นั้นมีอยู่มาก ซึ่งข้อดีของการสกัดข้อมูลด้วย DOM นั้น โดยส่วนมากงานวิจัยต่างๆ ที่นำเอาวิธีการนี้ไปใช้เป็นหลักจะสามารถสกัดข้อมูลได้แบบอัตโนมัติ เนื่องจากมีความยืดหยุ่นสูงในการเดินทางเข้าออกโหนดต่างๆ ในหน้าเว็บ ดังที่ได้ถูกกล่าวไว้ในงานวิจัยของ Sleiman, H. and Conchuelo [13]. แต่ข้อเสียของ DOM Model ก็คือ ยังขึ้นกับโครงสร้างของ HTML ดังนั้นจึงเป็นปัญหาเมื่อนักพัฒนาเว็บไซต์นั้นสร้างโครงสร้างที่ไม่ตรงตามมาตรฐาน จึงทำให้เป็นการยากที่จะทำให้วิธีของ DOM ยืดหยุ่นพอในการรองรับโครงสร้างของหน้าเว็บเพจได้หลากหลายรูปแบบ

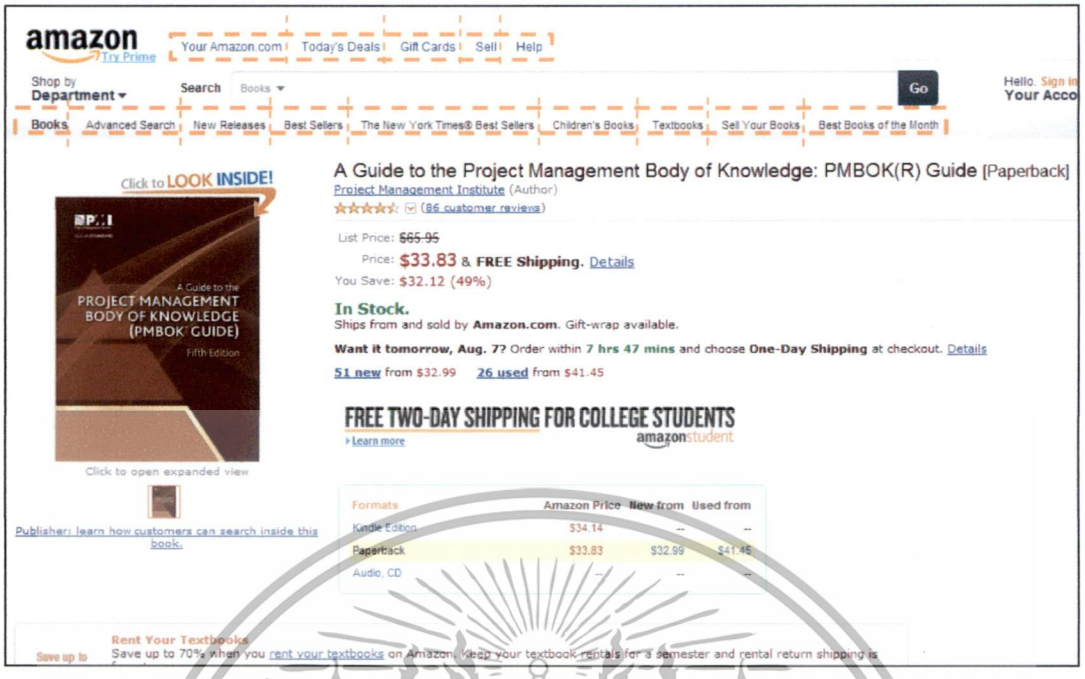
2.1.3 วิธีที่ขึ้นกับการมองเห็นของมนุษย์ (Vision Based)

วิธีที่ขึ้นกับการมองเห็น หรือ Vision Based ที่ถูกนำมาใช้งานทางด้านการสกัดข้อมูลจากหน้าเว็บเพจ คือการนำเอาสิ่งที่มองเห็นหรือโครงสร้างของหน้าเว็บเพจแต่ละเว็บออกมาแปลงโครงสร้างเพื่อนำไปใช้งานในด้านอื่นๆ ต่อไป โดยส่วนมากจะใช้ในการจัดกลุ่มขององค์ประกอบหรือตำแหน่งต่างๆ เพื่อแยกแยะประเภทของส่วนนั้นๆ (Web Page Segmentation) ดังงานวิจัยของ Cai, D., Yu, S., Wen, J. และ Ma, W. [3] ซึ่งโมเดลประเภท vision based จะแตกต่างกับโมเดลประเภทอื่นตรงที่ไม่ขึ้นกับโครงสร้างของ HTML ที่นักพัฒนาเว็บไซต์เขียนขึ้นมา เนื่องจากโมเดลประเภทนี้ใช้เทคนิคเรื่อง CSS Box Model ดังรูปที่ 2.7 มาช่วยในการตรวจหาและจัดกลุ่มเพื่อสร้างหน้าเว็บขึ้นมาเป็นโครงสร้างใหม่ โดยถูกเรียกว่า Wrapper โดย Wrapper จะทำหน้าที่เป็นเพียงแค่โครงสร้างแบบใหม่เท่านั้น ซึ่งเนื้อหาและรายละเอียดเดิมจะยังคงอยู่ (ทั้งนี้ขึ้นกับวัตถุประสงค์ของงานวิจัยนั้นๆ ว่าส่วนใดที่เป็นเป้าหมาย)



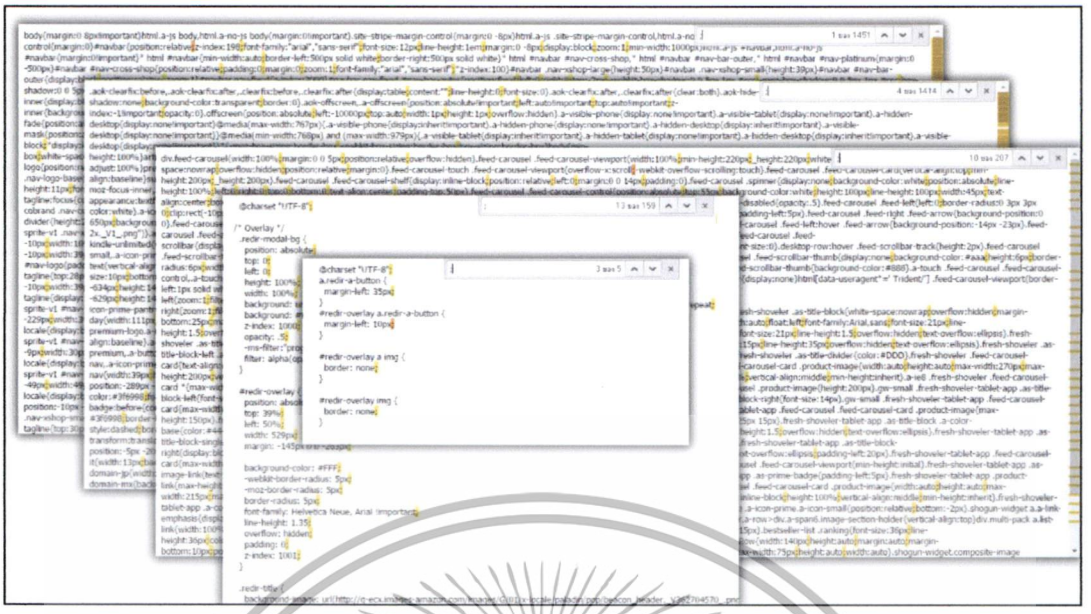
รูปที่ 2.7 ตัวอย่างโครงสร้างของ Box Model ในภาษาโปรแกรม CSS

การทำงานโดยทั่วไปของโมเดลประเภทนี้ โดยส่วนมากจะใช้ภาษาโปรแกรม CSS (Cascading Style Sheet) เข้ามาดำเนินการ เนื่องจากเป็นภาษาที่ใช้ในการตกแต่ง จัดวาง ระบุตำแหน่งของโหนดต่างๆ ในโครงสร้างของ HTML โดยในวิธีการนี้จะใช้ประโยชน์จากคุณลักษณะของแต่ละโหนดที่สำคัญ เช่น Position X, Position Y, Width, Height, Font-Size, Font-Weight เป็นต้น ซึ่งก็คือ ตำแหน่งระยะห่างจากขอบซ้ายบนสุดในแนวนอน ตำแหน่งระยะห่างจากขอบซ้ายบนสุดในแนวตั้ง ความกว้างของกล่อง ความสูงของกล่อง ขนาดตัวอักษร ความหนาของตัวอักษร ตามลำดับ ซึ่งคุณสมบัติเหล่านี้จะมีอยู่ในทุกๆ โหนดของ HTML Tags เพื่อทำให้การออกแบบออกมาในรูปแบบที่นักพัฒนาต้องการ โดยก้าวข้ามข้อจำกัดของเรื่องโครงสร้าง HTML ไป และด้วยคุณสมบัติเหล่านี้เป็นพื้นฐาน จึงทำให้วิธีเหล่านี้ใช้คุณสมบัติต่างๆเพื่อทำการจัดกลุ่มของโหนด HTML ยกตัวอย่างเช่น ถ้าหาก Position X มีค่าเท่ากัน และ Position Y มีค่าเพิ่มขึ้นทีละเท่าๆกัน แสดงว่าแท็กเหล่านั้นมีการเรียงตัวกันเป็นแนวนอน และหากแท็กเหล่านั้นมีขนาดความสูงเท่ากันหมด ก็จะมีความเป็นไปได้ที่จะจัดกลุ่มให้แท็กเหล่านั้นเป็นแท็กของเมนูในแนวนอนดังรูปที่ 2.8 และหากเพิ่มข้อจำกัดว่าทุกแท็กจะต้องเป็น หรือ มีแท็ก <a> อยู่ด้วย ก็จะเป็นประโยชน์ต่อการนำไปใช้งานในเรื่องการเก็บเกี่ยวลิ้งค์จากหน้าเว็บ เช่นการจัดทำแผนผังเว็บไซต์ (Site Map) หรืออาจประยุกต์ใช้งานกับเว็บบอท (Web Bot) เพื่อลดภาระในการเก็บเกี่ยวลิ้งค์ที่ไม่จำเป็นออกไป เป็นต้น



รูปที่ 2.8 ผลลัพธ์ตัวอย่างวิธีการจัดกลุ่มแบบ Vision Based

สำหรับข้อดีที่เป็นจุดเด่นของ Vision Based ดังที่กล่าวไปแล้ว คือ ไม่ขึ้นกับโครงสร้างเดิมของ HTML เนื่องจากทำงานด้วย Box Model จึงทำให้วิธีการประเภทนี้ได้เปรียบเนื่องจากมีความยืดหยุ่นสูงสามารถรองรับเว็บไซต์ได้หลากหลายรูปแบบ ส่วนข้อเสียของโมเดลประเภทนี้ก็คือ การทำงานนั้นค่อนข้างช้า เนื่องจากว่าต้องทำการอ่านเอกสารในการออกแบบ (CSS Source Code) ในทุกหน้า และทุกคุณสมบัติ ซึ่งเมื่อเจอเว็บไซต์ใหญ่ๆ จะทำงานช้ามาก เพราะมีคลาสที่ใช้ในการตกแต่ง (CSS Class) จำนวนมาก ซึ่งมีหลายคลาสไม่ได้ถูกใช้ในหน้าเว็บที่โมเดลต้องการ โดยปัญหาเหล่านี้อาจแก้ไขได้ด้วยการทำการจัดเตรียมข้อมูล (Pre-Processing) ของเอกสาร CSS ก่อนนำไปใช้เพื่อลดเวลาในการประมวลผล ยกตัวอย่างเช่น รูปที่ 2.9 เป็นตัวอย่าง Source Code ของเว็บไซต์ Amazon โดยมีไฟล์ CSS ทั้งหมด 4 ไฟล์ มีจำนวนคลาสรวมกันทั้งสิ้น 1,864 คลาส และมีคุณสมบัติ (Attribute) ทั้งหมด 3,236 คุณสมบัติ



รูปที่ 2.9 ตัวอย่าง Source Code ของไฟล์ CSS ในเว็บไซต์ Amazon

บทที่ 3
การดำเนินงานวิจัย

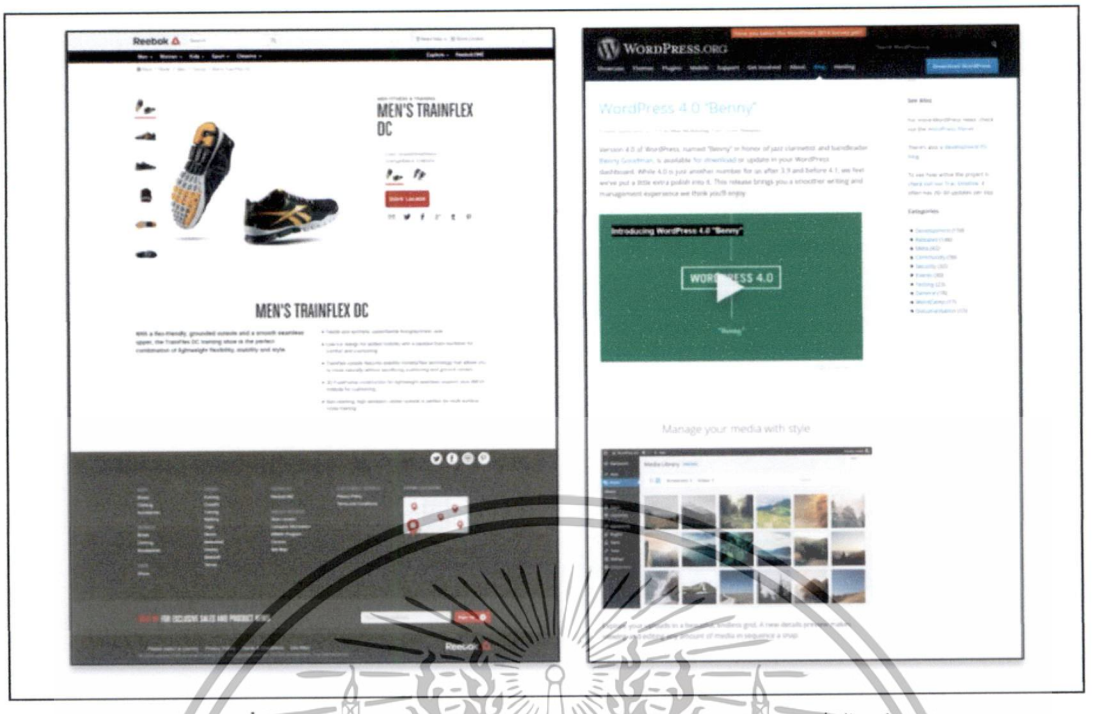
สำหรับงานวิจัยนี้ได้แบ่งการทำงานเป็น 3 ขั้นตอนใหญ่ ซึ่งประกอบด้วย การจัดเตรียมข้อมูล การตรวจจับโหนดหัวข้อ และการระบุโหนดที่มีข้อมูลอยู่อย่างหนาแน่นจากความหนาแน่นของโหนด โดยวิธีการนี้จะถูกเรียกว่า Subject Detective and Node Density (SDND)

3.1 การจัดเตรียมข้อมูล

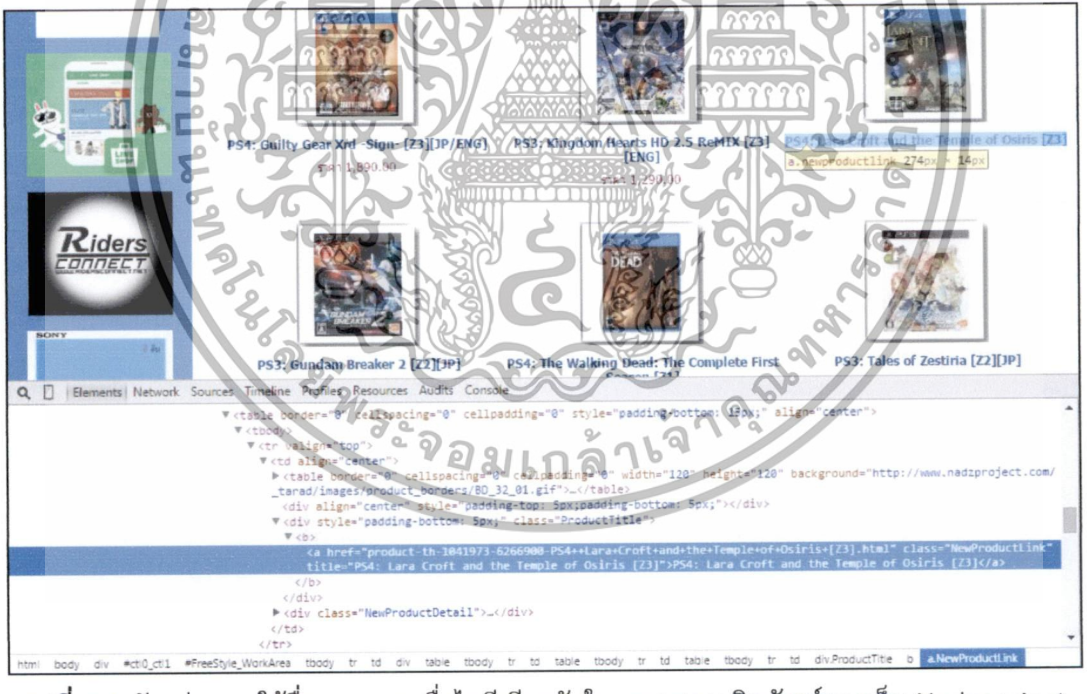
การจัดเตรียมข้อมูลก่อนการดำเนินงานวิจัย สำหรับงานวิจัยนี้มีความจำเป็นที่จะต้องจัดเตรียมข้อมูลก่อน เนื่องจากในส่วนของหน้าเว็บไซต์แต่ละหน้ามีความแตกต่างกันในเรื่องของการจัดวางตำแหน่งขององค์ประกอบต่างๆ บนหน้าเว็บ พฤติกรรมของนักพัฒนาเว็บไซต์ จุดมุ่งหมายของเว็บไซต์ และอื่นๆ จากรูปที่ 3.1 ทางด้านซ้ายเป็นตัวอย่างจากเว็บไซต์ร้านค้าออนไลน์ ส่วนทางด้านขวาเป็นตัวอย่างจากเว็บไซต์ประเภทแสดงบทความ ซึ่งมีความแตกต่างกันตรงที่ การจัดวาง โดยทางด้านซ้ายจะมีส่วนของเนื้อหาที่มีความกว้างเต็มจอ ส่วนทางด้านขวาจะมีเนื้อหากว้างโดยประมาณ 75 เปอร์เซ็นต์ของจอ และที่เหลือเป็นเมนูย่อย ส่วนของความแตกต่างกันระหว่างเว็บไซต์อีกส่วนคือมีการเรียกใช้ชื่อแท็กในการเข้าถึงไม่เหมือนกัน นอกเหนือจากนั้นยังมีความแตกต่างทางด้านองค์ประกอบภายในโหนดเนื้อหาที่ต้องการ หากลองสังเกตจะเห็นว่าในส่วนของเนื้อหาทางด้านซ้ายจะมีการจัดเรียงค่อนข้างซับซ้อนเมื่อแบ่งเป็นบล็อกเล็กๆ ส่วนทางด้านขวาจะมีการจัดวางเรียงตัวกันลงมาในลักษณะแนวตั้ง โดยสลับกันระหว่างย่อหน้าข้อความ วิดีโอ และรูปภาพ เป็นต้น หน้าที่ของกระบวนการนี้คือการเตรียมความพร้อมให้กับข้อมูลเพื่อนำไปใช้ในการทดลอง โดยกระบวนการจะมีดังนี้คือ เก็บลิ้งค์ข้อมูลของหน้าเว็บไซต์ที่ต้องการนำไปใช้ในการทดลอง คัดกรองข้อมูลที่ไม่จำเป็นต่อการทดลองออก โดยกระบวนการจัดเตรียมข้อมูลนี้จะทำงานเหมือนกันทุกโมเดลในการนำไปทดลอง เพื่อวัดประสิทธิภาพของแต่ละโมเดล

3.1.1 การเก็บลิ้งค์ข้อมูลของหน้าเว็บไซต์ที่ต้องการ

ข้อมูลสำหรับงานวิจัยนี้จะจัดทำโดยการสร้างซอฟต์แวร์สกัดข้อมูลจากหน้าเว็บเพจอย่างง่ายขึ้นมาเพื่อการเก็บลิ้งค์ของหน้าที่มีรายละเอียดอยู่อย่างหนาแน่น โดยระบุกฎไว้ว่าให้เก็บเฉพาะแท็ก <a> ซึ่งเป็นตัวแทนของลิ้งค์ทั้งหมดในหน้านั้นๆ และจะประกาศชื่อคลาส หรือชื่อของไอดีของลิ้งค์รายการที่ต้องการนั้นๆ ไว้ในแต่ละเว็บไซต์ เนื่องจากหนึ่งเว็บไซต์จะมีคลาสหรือไอดีที่เหมือนกันหมดในเว็บไซต์นั้นๆ ดังรูปที่ 3.2 หลังจากนั้นจึงรวบรวมโดยการเขียนลงไฟล์ในรูปแบบของไฟล์ CSV (Comma Separate Value) เพื่อให้ง่ายต่อการนำไปใช้งานต่อ ทำซ้ำอย่างนี้จนกระทั่งได้รับข้อมูลที่ต้องการครบถ้วนในทุกเว็บไซต์



รูปที่ 3.1 ตัวอย่างความแตกต่างในหลายด้านระหว่าง 2 เว็บไซต์



รูปที่ 3.2 ตัวอย่างการใช้ชื่อคลาสและชื่อไอดีเดียวกันในการแสดงผลิตภัณฑ์ของเว็บ Nadzproject

3.1.2 การคัดกรองข้อมูลที่ไม่จำเป็นต่อการทดลองออก

กระบวนการนี้จัดทำขึ้นเพื่อลดเวลาในการประมวลผลของการทดลอง และลดปริมาณของสิ่งรบกวน (Noise Node) ในหน้าเว็บออกเพื่อเพิ่มความแม่นยำในแต่ละโมเดลที่นำไปทดลอง ยกตัวอย่างเช่น ลิ้งค์ <a> ที่มีการเชื่อมโยงที่ไม่ใช่ URL ดังรูปที่ 3.3 และแท็กที่ไม่มีข้อความอยู่ภายในหรือ แท็กว่าง เนื่องจากแท็กว่างบางแท็กมีวัตถุประสงค์ในการสร้างขึ้นมาเพื่อทำการโต้ตอบกับผู้ใช้งานเท่านั้น จึงเป็นเหตุผลที่ว่าแท็กเหล่านี้จะถูกจัดให้เป็นสิ่งรบกวนต่อการทำงานของกระบวนการนี้อย่างชัดเจน

```
<a class="btnv6_green_white_innerfade btn_medium" href="javascript:addToCart( 40649);">
        <span>เพิ่มลงในรถเข็น</span>
</a>
```

รูปที่ 3.3 ตัวอย่างของลิ้งค์โหนดที่จะถูกตัดออกก่อนนำไปคำนวณ

3.2 การตรวจหาตำแหน่งของโหนดหัวข้อ (Subject Detection)

การตรวจหาตำแหน่งของหัวข้อหรือ Subject Detection นั้น จะถูกทำก่อนที่จะหาตำแหน่งของโหนดเนื้อหา เนื่องจากการสังเกตุที่พบว่าโหนดหัวข้อนั้นส่วนใหญ่จะอยู่ภายในโหนดเนื้อหา และโหนดหัวข้อนั้นสามารถที่จะสังเกตได้ง่ายกว่าส่วนอื่นๆ เนื่องจากนักพัฒนาเว็บไซต์ตั้งใจให้ส่งผลกระทบทางด้าน SEO (Search Engine Optimization) ยกตัวอย่างเช่น โหนดหัวข้อนั้น บ่อยครั้งจะเป็นแท็ก <H1> ดังรูปที่ 3.4 (ในกระบวนการ SEO <H1> จะถูกให้น้ำหนักมากที่สุดในการใช้เป็น คีย์เวิร์ด ของหน้าเว็บไซต์นั้นๆ) โหนดหัวข้อจะมีความคล้ายกับแท็ก Title และแท็ก Meta ในส่วนของคีย์เวิร์ด ดังรูปที่ 3.5 (ในกระบวนการ SEO แท็ก Title และ Keyword มีผลกระทบกับการถูกค้นหาจาก Search Engine) โดยการทำงานของโมเดลในการตรวจจับโหนดหัวข้อดังอัลกอริธึมที่ 1 และรูปที่ 3.6 มีดังนี้

3.2.1 ดาวน์โหลด HTML Source Code จากลิ้งค์ที่อยู่ของเว็บไซต์

ดาวน์โหลด HTML Source Code จากลิ้งค์ที่อยู่ของเว็บไซต์ที่เก็บมาในขั้นตอนการจัดเตรียมข้อมูล (ลิ้งค์ถูกเก็บอยู่ในรูปแบบของ CSV ไฟล์) หลังจากดาวน์โหลด Source Code เสร็จจะทำการตัดแท็กที่ไม่มีประโยชน์ออกไปตามที่ได้กล่าวไว้แล้วในกระบวนการจัดเตรียมข้อมูล

3.2.2 แปลง HTML Source Code เป็น DOM Tree

ทำการแปลงรูปแบบของ HTML Source Code ให้กลายเป็น DOM Tree เพื่อให้สามารถเข้าถึงและจัดการส่วนต่างๆ ของเอกสารได้ง่าย โดยการแปลงนั้นในงานวิจัยนี้จะใช้ HTML Tidy Node ซึ่งเป็นไลบรารี่ที่ได้รับความนิยมในการเขียนโปรแกรมในหลายภาษา

3.2.3 ตรวจหาแท็กที่มีความเป็นไปได้ที่จะเป็นโหนดหัวข้อ

ทำการตรวจหาแท็กที่มีความเป็นไปได้ที่จะเป็นโหนดหัวข้อทั้งหมดและทำสัญลักษณ์ไว้ เพื่อจะได้นำไปทำการคำนวณน้ำหนักต่อไป โดยในที่นี่จะให้ S = {<H1>, <H2>, <H3>, <DIV>, <SPAN>} (สำหรับการนำไปใช้งาน สามารถปรับได้ตามความเหมาะสม)

รูปที่ 3.4 ตัวอย่างของเว็บไซต์ Nadzproject ที่ใช้แท็ก <H1> เป็นหัวข้อ

H1 : <h1 class="title-32 vmargin8" itemprop="name">Men's Trainflex DC</h1>

Title : <title>Reebok Men's TrainFlex DC | Reebok International</title>

Keyword : <meta name="keywords" content="Men's TrainFlex DC"/>

รูปที่ 3.5 ตัวอย่างความคล้ายกันของแท็กหัวข้อ (H1) Title และ Keyword จากเว็บไซต์ Global Reebok

3.2.4 คำนวณค่าน้ำหนักมวลรวมของทุกๆ แท็ก

การคำนวณค่าน้ำหนักมวลรวมของทุกๆ แท็กในกลุ่มของแท็กที่มีความเป็นไปได้ที่จะเป็นโหนดหัวข้อจะใช้สูตรที่กำหนด ดังรูปที่ 3.7 โดยที่แต่ละสูตรจะมีการปรับลดค่าพารามิเตอร์เพื่อให้อยู่ในช่วง 0 ถึง 1 เพื่อความง่ายในการนำไปคำนวณเพื่อเปรียบเทียบ โดยการคำนวณแต่ละผลลัพธ์ค่าน้ำหนักจะถูกกำหนดไว้แล้วตั้งแต่แรก ในงานวิจัยนี้จะกำหนดค่าน้ำหนักที่สามารถปรับลดได้

(Adjustable Weight) ดังรูปที่ 3.7 โดยการกำหนดค่าน้ำหนักเป็นการกำหนดค่าจากความเป็นไปได้ของการเป็นโหนดหัวข้อตามลำดับโดยผ่านการทดลองและเลือกค่าที่ดีที่สุด (สามารถดูผลการทดลองเพิ่มเติมได้ที่ภาคผนวก) การใช้สูตรต่างๆ เพื่อช่วยคำนวณมาจากข้อสังเกตที่ได้กล่าวไปแล้วโดยจะถูกเขียนอยู่ในรูปสมการการหาน้ำหนักมวลรวมของความเป็นไปได้ที่จะเป็นโหนดหัวข้อ ซึ่งใช้สำหรับเปรียบเทียบน้ำหนักกับทุกแท็กสมาชิกที่ *i* (*<H1>, <H2>, <H3>, <DIV>, <SPAN>*) ดังสมการ 3.1

$$W_i = (\alpha \times N_i) + (\beta \times T_i) + (\gamma \times K_i) + (\delta \times C_i) \quad (3.1)$$

โดยที่ W_i คือ ค่าน้ำหนักมวลรวม

N_i คือ ชื่อของแท็กสมาชิกที่ *i*

T_i คือ เนื้อหาของ *Title* แท็ก

K_i คือ เนื้อหาที่อยู่ภายในแท็ก *<META>* ที่มีชื่อว่า keyword

C_i คือ ค่าคุณสมบัติทาง CSS

ส่วนค่าคงที่ α, β, γ และ δ ใช้เพื่อถ่วงน้ำหนักโดยมีผลรวมเท่ากับ 1 (สามารถปรับเพื่อเพิ่มลดตามความเหมาะสมได้)

จากสมการที่ 3.1 เราสามารถคำนวณหาค่าน้ำหนักของชื่อแท็กได้จากสมการ 3.2

$$N_i = \begin{cases} 1.00 & if\ name=<h1> \\ 0.75 & if\ name=<h2> \\ 0.50 & if\ name=<h3> \\ 0.25 & if\ name=<div>,<span> \end{cases} \quad (3.2)$$

โดยที่ N_i คือ ค่าน้ำหนักของชื่อแท็ก HTML ดังที่กล่าวไปแล้วโดยกำหนดค่าให้กับชื่อแท็กดังสมการที่ 3.2 โดยน้ำหนักของ *<h1> <h2> <h3> * และ ** มีค่าเท่ากับ 1, 0.75, 0.5 และ 0.25 ตามลำดับ

ค่าความเหมือนกันระหว่างเนื้อหาในแท็กเป้าหมายกับแท็ก Title สามารถคำนวณได้ดังสมการที่ 3.3

$$T_i = \frac{|A \cap B_i|}{max(|A|,|B_i|)} \quad (3.3)$$

โดยที่ T_i คือ ความเหมือนกันระหว่างแท็ก *Title* และแท็กสมาชิกที่ *i*

A คือ เซตของคำในแท็ก *Title*

B_i คือ เซตของคำทั้งหมดในแท็กสมาชิกที่ *i*

$max(|A|, |B_i|)$ คือค่าที่มากที่สุดระหว่างจำนวนสมาชิกของ A และ B_i

ค่าความเหมือนกันระหว่างแท็กเป้าหมายและแท็ก Keyword สามารถคำนวณได้ดังสมการที่ 3.4

$$K_i = \frac{|C \cap D_i|}{max(|C|, |D_i|)} \tag{3.4}$$

โดยที่ K_i คือ ผลลัพธ์จากการวัดความคล้ายกันของคำระหว่างแท็ก *Keyword* และแท็กสมาชิกที่ i โดยการคำนวณจะเหมือนกันกับสมการที่ 3.3 เพียงแต่เปลี่ยนจาก A เป็น C และ B เป็น C โดย C คือ คำที่อยู่ในแท็ก *Keyword* แทน

ในส่วนของค่าคุณสมบัติของ CSS สามารถคำนวณได้ดังสมการที่ 3.5

$$C_i = (\kappa \times display_i) + (\lambda \times fontweight_i) + (\mu \times fontsize_i) \tag{3.5}$$

โดยที่ C_i คือ ผลลัพธ์จากการคำนวณค่าคุณสมบัติของ CSS (CSS Properties) ในสมการนี้จะประกอบด้วย 3 ค่าโดยทั้ง 3 ค่านี้จะได้มาจากเอกสาร *CSS* ในแต่ละหน้าเว็บ (*<stylesheet>*) ได้แก่ ค่าคุณลักษณะการจัดวางแท็ก (Display Property) ซึ่งสามารถคำนวณได้ดังสมการที่ 3.6

$$display_i = \begin{cases} 1 & if\ display = block \\ 0 & if\ otherwise \end{cases} \tag{3.6}$$

โดยที่ $display_i$ คือ รูปแบบในการแสดงแท็กของแท็กที่ *i* โดย *display = block* คือในบรรทัดนั้นจะแสดงเพียงแท็กเดียว (ขึ้นกับแท็กแม่) ซึ่งในงานวิจัยนี้จะกำหนดค่าเป็น 1 และหากเป็นค่าอื่นจะให้ค่าเป็น 0

ค่าความหนาของตัวอักษร (Font Weight) สามารถคำนวณได้ดังสมการที่ 3.7

$$fontweight_i = \begin{cases} 1 & if\ fontweight = bold, bolder \\ 0 & if\ otherwise \end{cases} \tag{3.7}$$

โดยที่ $fontweight_i$ จะมีค่าเป็น 1 ก็ต่อเมื่อมีค่าเท่ากับ *bold* หรือ *bolder* ซึ่งหมายถึงตัวอักษรหนา โดยในกรณีอื่นๆ จะมีค่าเป็น 0 และค่าขนาดของตัวอักษร (Font Weight) โดยจะเปรียบเทียบกับขนาดของตัวอักษรที่ใหญ่ที่สุด และเล็กที่สุดในหน้านั้นๆ ซึ่งจะเปรียบเทียบโดยใช้วิธีการแปลงให้อยู่ใน

หน่วยของ em ก่อนการเปรียบเทียบ และทำการลดรูปโดยใช้วิธีการ Min Max Normalization ค่าขนาดของตัวอักษรสามารถคำนวณได้ดังสมการที่ 3.8

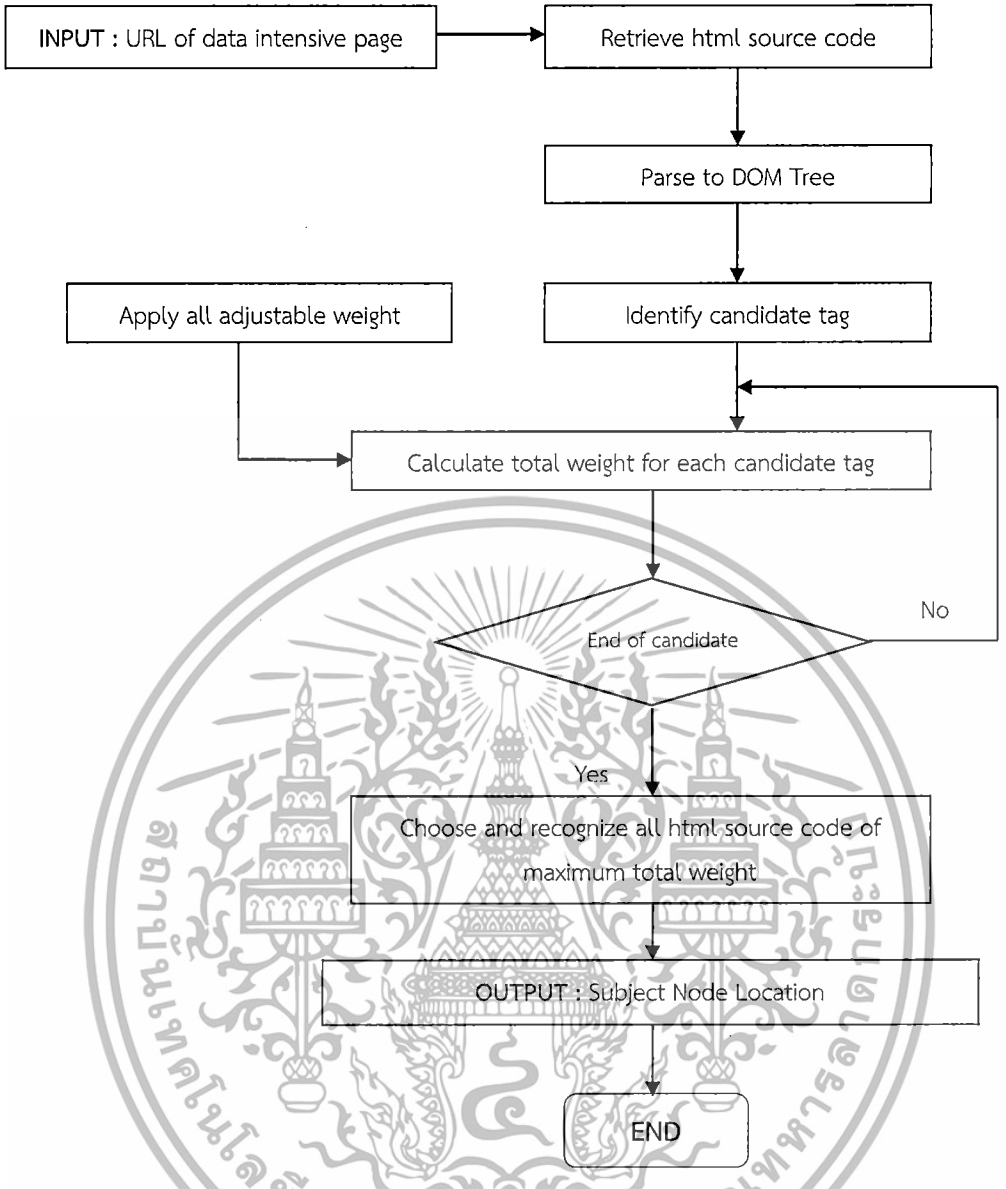
$$fontsize_i = \frac{(fs_i - minfs)}{(maxfs - minfs)} \quad (3.8)$$

โดยที่ $fontsize_i$ คือ ผลลัพธ์จากการคำนวณน้ำหนักของขนาดตัวอักษรในแท็กที่ i

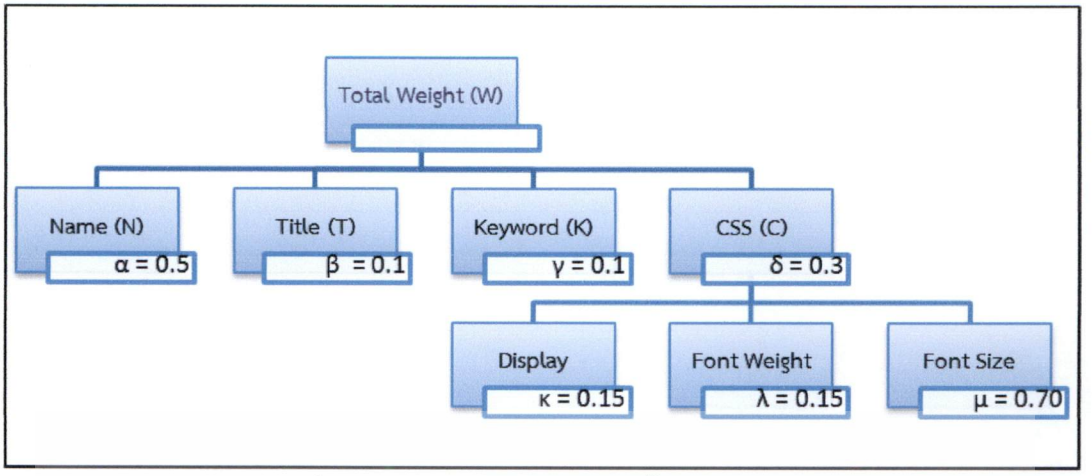
fs_i คือขนาดตัวอักษรในแท็กที่ i

$maxfs$ คือ ขนาดตัวอักษรที่ใหญ่ที่สุดในหน้าเว็บนั้น

$minfs$ คือ ขนาดตัวอักอักษรที่เล็กที่สุดในหน้าเว็บนั้น โดยจะแปลงขนาดตัวอักษรเป็นหน่วย *em* ก่อนการคำนวณ



รูปที่ 3.6 กระบวนการทำงานของโมเดลการหาตำแหน่งโหนดหัวข้อ (Subject Detection)



รูปที่ 3.7 โครงสร้างการคำนวณค่าน้ำหนักมวลรวม

3.2.5 เลือกแท็กที่มีค่าน้ำหนักมวลรวมมากที่สุดเป็นโหนดหัวข้อ

หลังจากที่ได้ค่าน้ำหนักมวลรวมของทุกแท็กแล้วจึงทำการเลือกแท็กที่มีค่าน้ำหนักมวลรวมมากที่สุดเป็นโหนดหัวข้อ เพื่อนำไปใช้ในการสกัดข้อมูลในกระบวนการถัดไป

Algorithm 1: Subject Detection
Input: URL of data intensive page
Output: Subject node

1. Retrieve HTML Source Code
2. Parse HTML to DOM Tree
3. **for each** node *i* in DOM Tree **do**
4. **if** the tag name is an element of candidate tags **then**
5. Calculate total weight of node *i*
6. Find the node that has the maximum value of total weights and assign it to be a subject node.

3.3 การหาตำแหน่งของโหนดที่มีรายละเอียดอยู่อย่างหนาแน่น (Data Rich Region)

การหาตำแหน่งของโหนดที่มีรายละเอียดอยู่อย่างหนาแน่น คือ การหาว่าโหนดใดควรจะเป็นโหนดที่ถูกเลือกให้เป็นโหนดข้อมูลซึ่งดูจากอัตราการเพิ่มขึ้นระหว่างโหนดปกติและลิ้งค์โหนด โดยเริ่มการทำงานจากโหนดหัวข้อที่ได้รับมาจากขั้นตอนที่แล้ว และทำการถอยกลับไปยังโหนดแม่ของโหนดปัจจุบันทุกครั้งที่อัตราการเพิ่มขึ้นของลิ้งค์โหนดยังเป็นปกติ แต่หากเกิดความผิดปกติขึ้น จะหยุดการทำงานและกำหนดให้โหนดปัจจุบันนั้นเป็นโหนดเนื้อหาทันที ดังอัลกอริธึมที่ 2 และรูปที่ 3.8 โดยการทำงานจะมีดังนี้

3.3.1 รับตำแหน่งของโหนดหัวข้อจากขั้นตอนก่อนหน้า

รับตำแหน่งของโหนดหัวข้อจากขั้นตอนก่อนหน้าเพื่อนำมาคำนวณหาตำแหน่งของโหนดที่มีรายละเอียดอยู่อย่างหนาแน่น โดยการเริ่มต้นการทำงานจะเริ่มที่ตำแหน่งของโหนดหัวข้อที่ได้รับมา

3.3.2 หาค่าอัตราการเพิ่มขึ้นระหว่างโหนดปกติและลิ้งค์โหนด

คำนวณหาค่าอัตราการเพิ่มขึ้นระหว่างโหนดปกติและลิ้งค์โหนด โดยที่ค่า Threshold เก่าจะถูกกำหนดค่าเริ่มต้นเป็น 0 และหากการหารนั้นมีตัวหารเป็น 0 จะระบุผลลัพธ์เป็น 1 เพื่อป้องกันข้อผิดพลาดของโปรแกรมในขณะที่กำลังทำงาน การคำนวณค่า threshold สามารถคำนวณได้ดังสมการที่ 3.9

$$threshold = \frac{node\ density - link\ density}{link\ density} \tag{3.9}$$

โดยที่ *threshold* คือ อัตราการเพิ่มขึ้นของปริมาณโหนดปกติ (*node density*) และลิ้งค์โหนด (*link density*) เพื่อระบุว่าเมื่อไหร่จะหยุดทำงาน โดยที่หากความหนาแน่นของลิ้งค์โหนดเป็น 0 จะบังคับให้เป็น 1

3.3.3 เปรียบเทียบ Threshold โหนดก่อนหน้าและโหนดปัจจุบัน

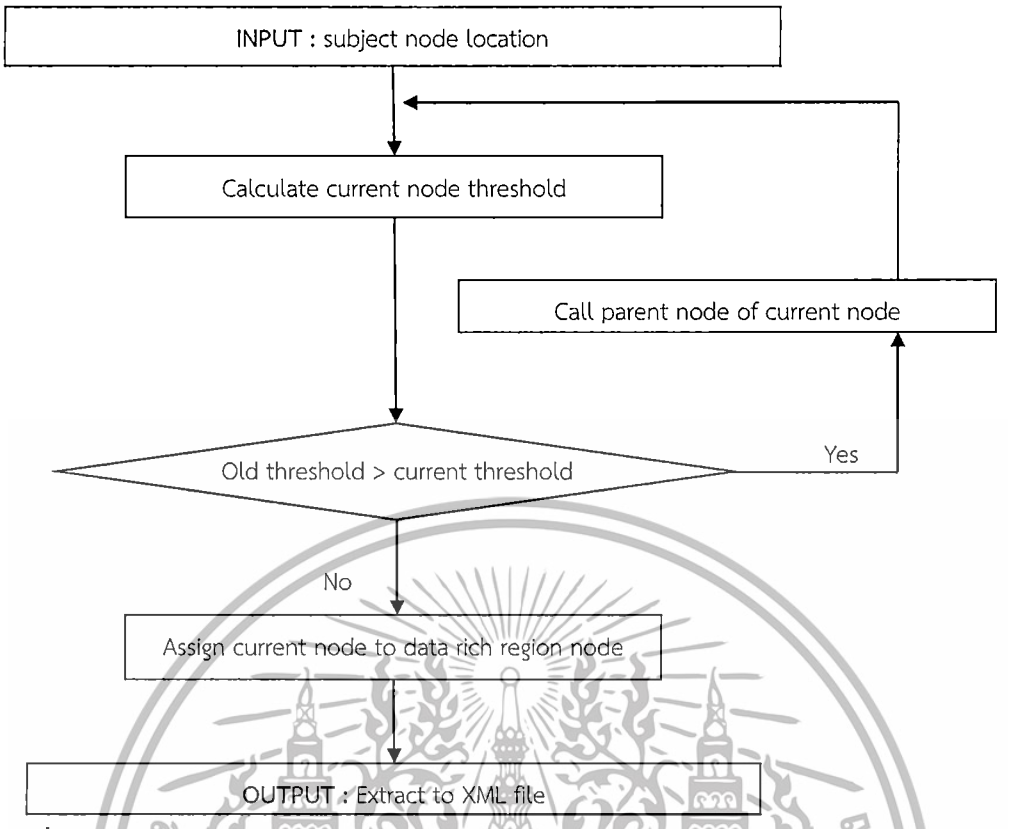
ทำการเปรียบเทียบระหว่าง Threshold ของโหนดก่อนหน้าและโหนดปัจจุบันเพื่อตรวจสอบเงื่อนไขการหยุดทำงานของโมเดล

3.3.3.1 หากผลการตรวจสอบออกมาว่า Threshold ใหม่มีค่ามากกว่า Threshold เก่า โมเดลจะทำการเรียกโหนดแม่ออกมาเป็นโหนดปัจจุบัน และกลับไปทำในข้อ 3.3.2 ใหม่

3.3.3.2 หากผลการตรวจสอบออกมาว่า Threshold เก่ามีค่ามากกว่า Threshold ใหม่ โมเดลจะทำงานต่อในขั้นตอนถัดไป คือ ข้อ 3.3.4

3.3.4 ทำการระบุโหนดก่อนหน้าของโหนดปัจจุบันให้เป็นโหนดข้อมูล

ทำการระบุโหนดก่อนหน้าของโหนดปัจจุบันให้เป็นโหนดข้อมูลหรือโหนดเนื้อหา (Data Rich Region Node) หลังจากนั้นจึงจบการทำงานโดยที่ผลลัพธ์จะเป็นโครงสร้างของภาษาเอชทีเอ็มแอลในรูปแบบ DOM



รูปที่ 3.8 กระบวนการทำงานของโมเดล การหาโหนดข้อมูลจากความหนาแน่นของโหนด (Node Density)

Algorithm 2: Node Density
Input: Subject node
Output: XML file

1. Set the current node equal to the subject node and set the previous threshold as 0
2. Compute the current threshold using equation 8
3. **if** current threshold ≥ previous threshold **then**
4. Assign previous threshold equal to the current threshold
5. Go to the parent node of the current node and assign it to be the current node
6. Go to step 3
7. **else**
8. Assign the current node to be the data rich region node
9. Extract the detail to XML file

3.4 การสกัดข้อมูล (Extraction)

การสกัดข้อมูลเป็นกระบวนการหลักสำหรับงานวิจัยนี้ โดยที่ในการใช้งานนั้นสามารถปรับแต่งผลลัพธ์หรือวิธีการแสดงข้อมูลได้ตามความเหมาะสม โดยในงานวิจัยนี้จะใช้ ภาษา XML (Extensible Markup Language) เป็นผลลัพธ์ในการแสดงผลออกมาเพื่อให้เข้าใจง่าย โดยมีแท็กต่างๆ ดังรูปที่ 3.9 ซึ่งหลังจากนั้นจะใช้ ภาษาโปรแกรม XSLT (Extensible Stylesheet Language Transformations) และ XPATH (XML Path Language) เพื่อเป็นภาษาในการเรียกค้นเพื่อตรวจสอบผลลัพธ์ระหว่างสิ่งที่ถูกสกัดออกมากับสิ่งที่ต้องการว่าตรงกันหรือไม่ในการตรวจสอบผลลัพธ์ และอีกเครื่องมือหนึ่งที่ใช้ก็คือเว็บบราวเซอร์ Google Chrome เนื่องจากมีระบบค้นหาที่สามารถเน้นสีคำที่ตรงกับคำค้นได้ และสามารถเปิดไฟล์ XML โดยผ่านการจัดเรียงให้ดูง่าย

รูปที่ 3.9 ตัวอย่างการแสดงผลบนบราวเซอร์ Chrome จากไฟล์ XML

บทที่ 4
ผลการวิจัยและการอภิปรายผล

4.1 ข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่ใช้ในการทดลองสำหรับงานวิจัยนี้ มีวัตถุประสงค์เพื่อพิสูจน์ความแม่นยำ และความยืดหยุ่นของโมเดลเป็นหลัก เพื่อให้รองรับกับหน้าเว็บไซต์ได้หลากหลายรูปแบบที่มีความแตกต่างกันทั้งทางด้านโครงสร้างและภาษา และสามารถนำไปประยุกต์ใช้จริงได้ในอนาคต ดังนั้น งานวิจัยนี้จึงทำการทดลองในลักษณะการนำไปใช้ในโปรแกรมจริง มีการเขียนโปรแกรมเพิ่มเติมในส่วนนอกเหนือจากทฤษฎีที่นำเสนอ ได้แก่ส่วนของการจัดเตรียมข้อมูล และส่วนของการสกัดข้อมูล ซึ่งส่วนเพิ่มเติมเหล่านี้สามารถปรับแต่งตามความต้องการของผู้นำไปใช้ได้ ยกตัวอย่างเช่น งานวิจัยนี้ใช้ XML เป็นผลลัพธ์ในการแสดงผล ผู้ที่นำไปใช้งานสามารถเลือกวิธีการเก็บข้อมูลเป็นประเภทอื่นได้ เช่น ฐานข้อมูลเชิงสัมพันธ์ (Relational Database) และในการแสดงผลลัพธ์ยังสามารถจัดเรียงหรือแสดงในรูปแบบเดิมได้ด้วย เนื่องจากโมเดลนี้สามารถคงรูปของโครงสร้างเดิมที่สกัดไว้ได้ เพราะการสกัดข้อมูลของโมเดลนี้จะได้ผลลัพธ์เพียงแค่โหนดเดียว แต่ตัวอย่างของผลลัพธ์ที่แสดงในงานวิจัยนี้ ได้ทำการดึงข้อความออกจากทุกโหนดลงในรูปแบบของตัวอักษรเรียบร้อยแล้ว ซึ่งกระบวนการหรือวิธีการในการเก็บข้อมูลนั้นได้ถูกกล่าวไว้ในบทที่ 3 ข้อ 3.1 ในส่วนของกระบวนการจัดเตรียมข้อมูล

4.1.1 เว็บไซต์ที่ใช้ในการทดลอง

ข้อมูลหรือเว็บไซต์ที่ใช้ในการทดลองสำหรับวิจัยนี้นั้น ดังที่กล่าวไปแล้วว่าต้องการความหลากหลายในตัวข้อมูล เนื่องจากเป็นวัตถุประสงค์หลักของงานวิจัยทางด้านนี้ที่ต้องการความยืดหยุ่นของโมเดล ดังนั้นการตัดสินใจเลือกใช้ข้อมูลในการทดลองมีความสำคัญมาก สำหรับงานวิจัยนี้ได้เลือกใช้ข้อมูลในการทดลองโดยพิจารณาดังตารางที่ 4.1

ตารางที่ 4.1 รายละเอียดของเว็บไซต์ที่ถูกใช้ในการทดลอง

Dataset URL	Extracted pages	Language	Size	Objective	Product type	Actual shop
http://www.hngcase.com/	150	Thai	Small	B2C	Mobile Case	No
http://www.nadzproject.com/?lang=th	150	Thai	Medium	B2C	Game	Yes
http://www.ksscom.co.th/	150	Thai	Medium	B2C	IT product	Yes
http://global.reebok.com/	150	English	Large	B2C	Reebok's brand	Yes
http://www.olx.co.th/?from=www	150	Thai	Large	C2C	All	No
http://www.asos.com/men/	150	English	Large	B2C	Fashion	No
https://www.jib.co.th/web/	150	Thai	Medium	B2C	IT product	Yes
http://www.360tvshopping.com/	147	Thai	Small	B2C	Fitness equipment	No

จากตารางที่ 4.1 รายละเอียดของเว็บไซต์ที่ใช้ในการทดลองได้แก่ ที่อยู่เว็บไซต์ของข้อมูลที่ใช้ (Dataset URL) จำนวนหน้าเว็บไซต์ที่นำมาทดลองของแต่ละเว็บไซต์ (Extracted Pages) ภาษาหลักที่ใช้ในเว็บไซต์ (Language) ขนาดของข้อมูลภายในเว็บไซต์ (Size) โดยส่วนของขนาดจะวัดจากปริมาณของจำนวนการเข้าชมเว็บไซต์เป็นหลัก วัตถุประสงค์ของเว็บไซต์ (Objective) ประเภทของสินค้าที่ขาย (Product Type) และการมีหน้าร้านสำหรับวางขายสินค้าตั้งอยู่จริง (Actual Shop) ซึ่งส่วนนี้จะเป็นส่วนประกอบของความน่าเชื่อถือของร้านค้าหรือหน้าเว็บไซต์นั้นๆ โดยที่ B2C (Business to Customer) คือ เว็บไซต์ประเภทผู้ขายเป็นฝ่ายของเจ้าของเว็บไซต์ และเปิดเว็บไซต์เพื่อขายของของตนเองเป็นหลัก และ C2C (Customer to Customer) คือ เว็บไซต์ประเภทที่เจ้าของเว็บไซต์ตั้งใจเปิดเว็บไซต์เพื่อเป็นเสมือนตลาดกลางในการ ซื้อ-ขายสินค้า โดยที่การซื้อ-ขายนั้น ทางเจ้าของเว็บไซต์จะไม่ร่วมรับผิดชอบกับการซื้อ-ขายระหว่างลูกค้าด้วยกันเอง ดังนั้นผู้ขายบนเว็บไซต์ประเภทนี้จึงมีความหน้าเชื่อถือค่อนข้างต่ำ และมีสินค้ามือสองขายอยู่มาก โดยรวมแล้วตารางที่ 4.1 มีวัตถุประสงค์เพื่อพิสูจน์ความน่าเชื่อถือของเว็บไซต์ต่างๆ ที่ถูกนำมาใช้ในการทดลองและพิสูจน์ถึงความยืดหยุ่นของโมเดลที่ต้องรองรับกับความหลากหลายทางโครงสร้างบนหน้าเว็บไซต์

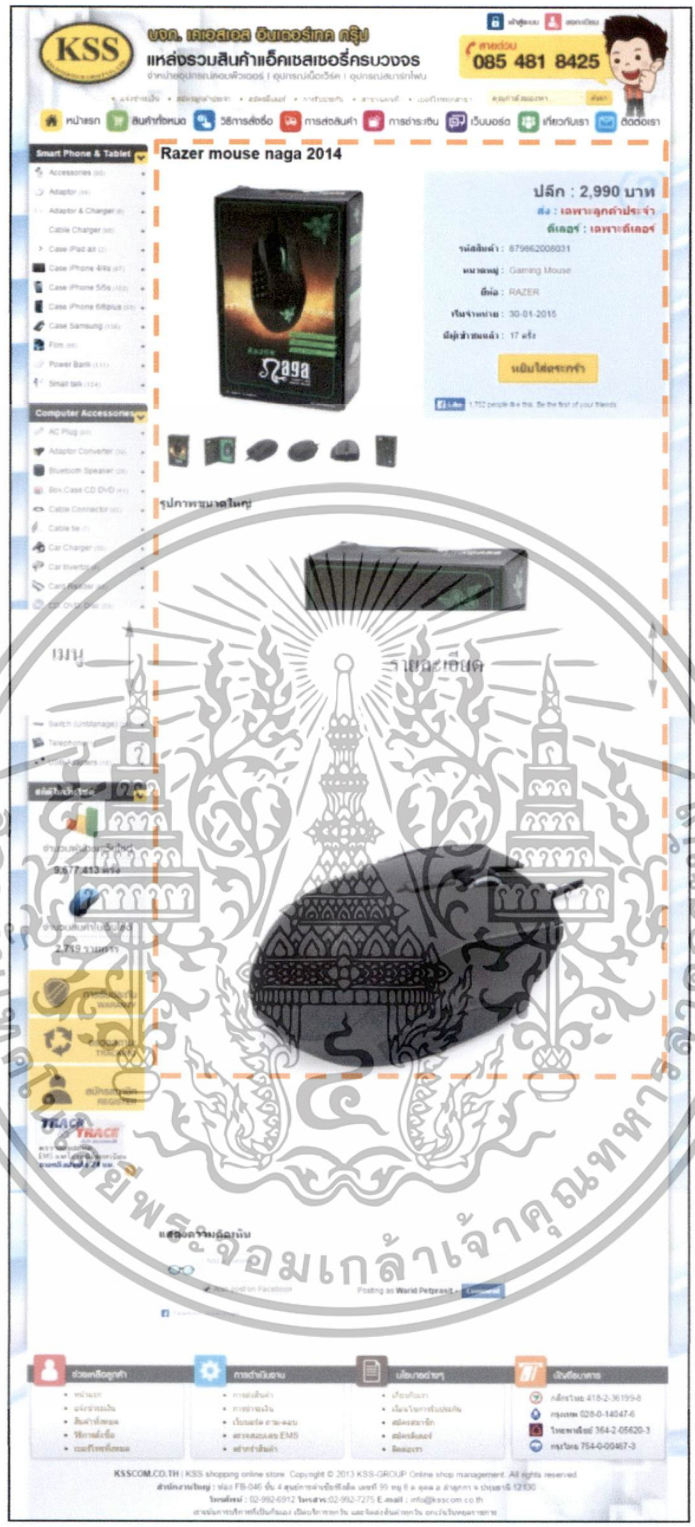
4.1.2 ตัวอย่างของหน้าเว็บไซต์ที่ใช้ในการทดลอง

รูปที่ 4.1 ตัวอย่างส่วนรายละเอียดผลิตภัณฑ์หน้าเว็บไซต์ HNGCASE

จากรูปที่ 4.1 แสดงถึงหน้ารายละเอียดผลิตภัณฑ์ของเว็บไซต์ HNGCASE ซึ่งเป็นเว็บไซต์ขายกรอบโทรศัพท์มือถือ ที่ได้รับความนิยมในระดับกลางซึ่งถูกสร้างขึ้นจากเว็บไซต์ที่ให้บริการสร้างเว็บไซต์ขายของฟรี (www.lnwshop.com)

รูปที่ 4.2 ตัวอย่างส่วนรายละเอียดผลิตภัณฑ์หน้าเว็บไซต์ Nadzproject

รูปที่ 4.2 คือหน้าเว็บไซต์ตัวอย่างของร้าน Nadz จากเว็บไซต์ www.nadzproject.com ที่มีหน้าร้านอยู่จริงหลายแห่ง และได้รับความนิยมในกลุ่มของผู้เล่นเกมคอนโซล

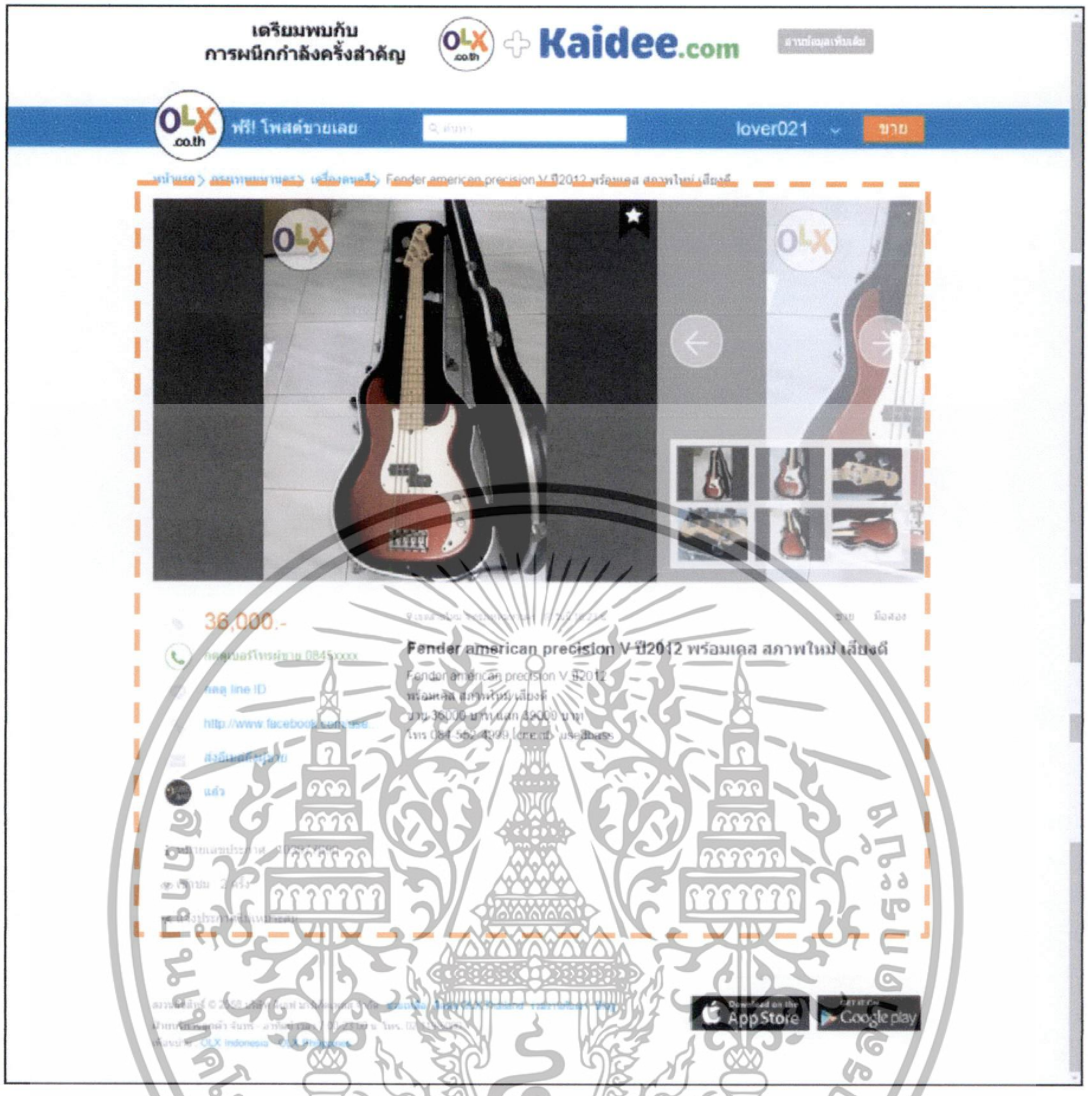


รูปที่ 4.3 ตัวอย่างส่วนรายละเอียดผลิตภัณฑ์หน้าเว็บไซต์ KSSCOM

รูป 4.3 คือหน้าเว็บตัวอย่างจากเว็บไซต์ ksscom ซึ่งได้รับความน่าเชื่อถือและมีอยู่หลายสาขาในประเทศไทย ขายสินค้าที่เกี่ยวกับอุปกรณ์ และอะไหล่คอมพิวเตอร์

รูปที่ 4.4 ตัวอย่างส่วนรายละเอียดผลิตภัณฑ์หน้าเว็บไซต์ Global Reebok

รูปที่ 4.4 คือหน้าเว็บตัวอย่างจากเว็บไซต์ global.reebok.com ซึ่งเป็นเว็บไซต์ศูนย์กลางของผลิตภัณฑ์ตราสินค้าที่ชื่อว่า reebok โดยที่เว็บไซต์นี้จะไม่มีการขายสินค้าผ่านเว็บไซต์ แต่จะเป็นเว็บไซต์สำหรับรวมรวบข้อมูลผลิตภัณฑ์ทั้งหมดของตราสินค้านี้ รวมถึงระบุแหล่งที่ตั้งของตัวแทนจำหน่ายทั่วโลก

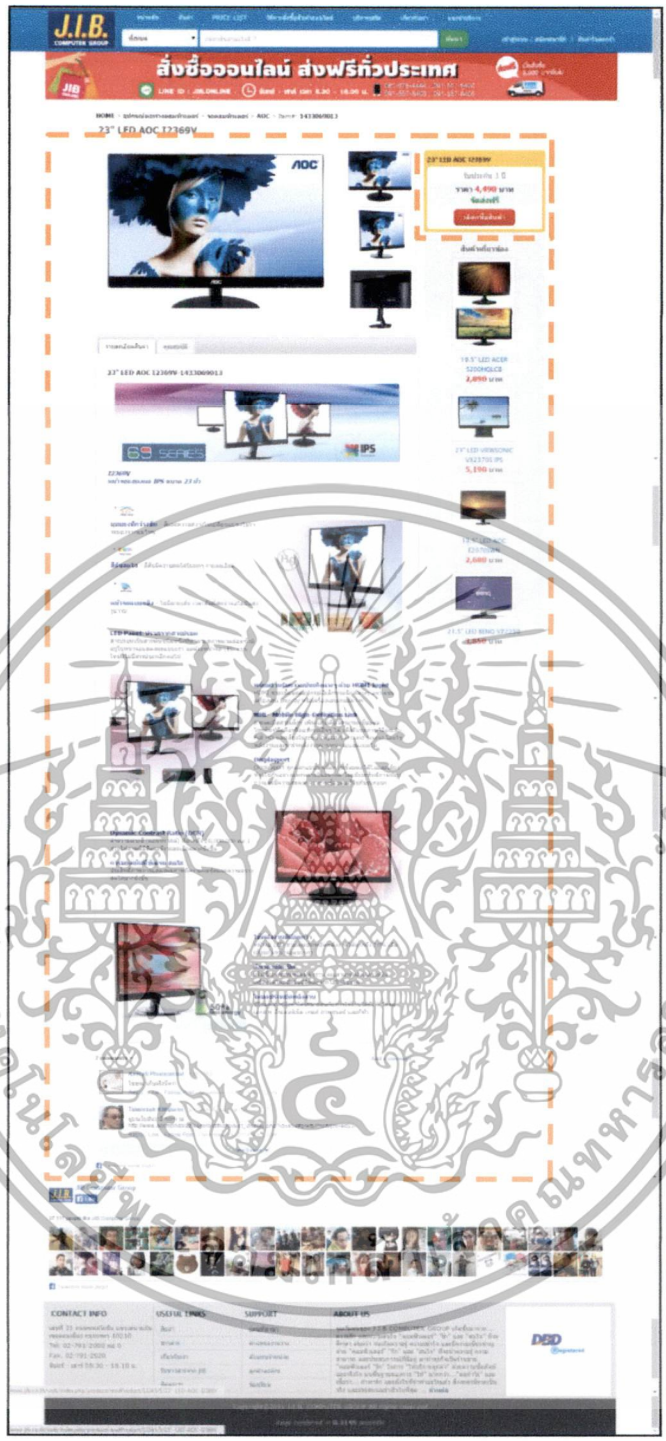


รูปที่ 4.5 ตัวอย่างส่วนรายละเอียดผลิตภัณฑ์หน้าเว็บไซต์ OLX

รูปที่ 4.5 คือรูปตัวอย่างเว็บไซต์ในหน้ารายละเอียดผลิตภัณฑ์ของเว็บไซต์ OLX ซึ่งเป็นเว็บไซต์ประเภทตลาดกลาง (customer to customer) โดยที่เว็บไซต์ OLX มีการโฆษณาผ่านทางโทรทัศน์ค่อนข้างบ่อยจึงทำให้ผู้ใช้งานทั่วไปรู้สึกคุ้นเคย และมีความน่าเชื่อถือ

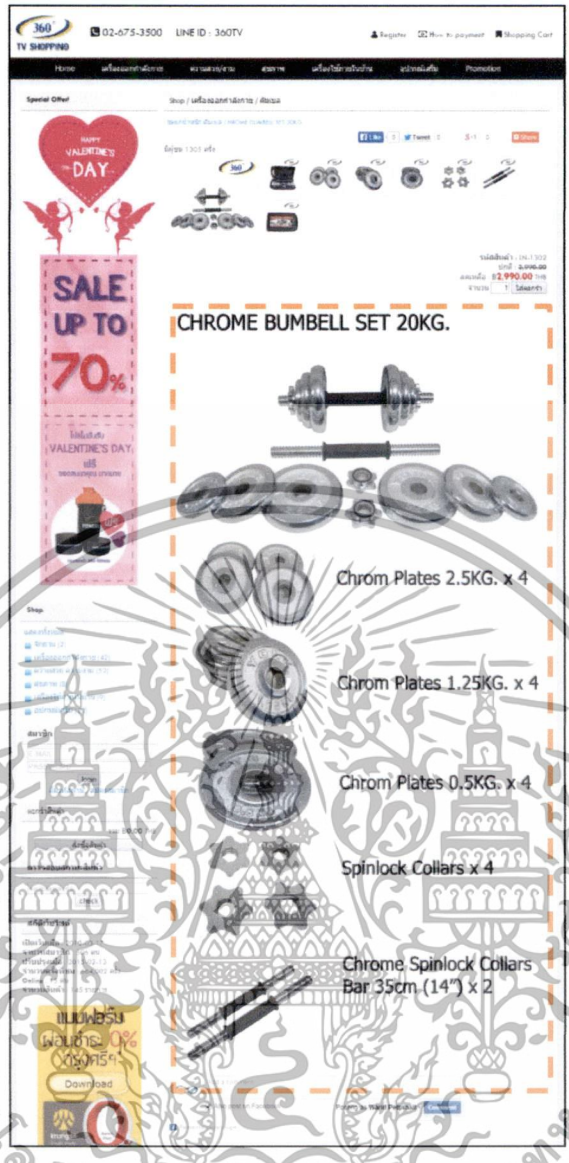
รูปที่ 4.6 ตัวอย่างส่วนรายละเอียดผลิตภัณฑ์หน้าเว็บไซต์ ASOS

รูปที่ 4.6 คือตัวอย่างของหน้าเว็บไซต์รายละเอียดผลิตภัณฑ์จากเว็บไซต์ ASOS ซึ่งเป็นเว็บไซต์ต่างประเทศและรองรับหลายภาษา โดยที่เว็บไซต์นี้จะถูกนำมาวัดความสามารถของงานวิจัยนี้ว่างานวิจัยนี้สามารถทำงานได้แม้จะต่างภาษากัน



รูปที่ 4.7 ตัวอย่างส่วนรายละเอียดผลิตภัณฑ์หน้าเว็บไซต์ JIB Computer

รูปที่ 4.7 คือตัวอย่างของหน้าเว็บไซต์ของร้าน JIB Computer ซึ่งได้รับความนิยมในกลุ่มของผู้ใช้งานคอมพิวเตอร์ทั่วไป จากรูป ในกรอบใหญ่ที่เป็นเส้นประคือส่วนที่โมเดลสามารถสกัดได้ ซึ่งมีสิ่งรบกวนมาด้วย คือส่วนของการแนะนำผลิตภัณฑ์ โดยส่วนนี้จะเป็นข้อจำกัดของโมเดลที่ได้อธิบายไว้ในส่วนท้ายของบทที่ 4



รูปที่ 4.8 ตัวอย่างส่วนรายละเอียดผลิตภัณฑ์หน้าเว็บไซต์ TV360Shopping

รูปที่ 4.8 ตัวอย่างหน้าเว็บของเว็บไซต์ TV360Shopping ซึ่งถูกโฆษณาบ่อยผ่านทางโทรทัศน์และมีหน้าเว็บไซต์สำหรับขายสินค้าออนไลน์ ซึ่งเป็นที่นิยมในการสั่งซื้อสินค้าผ่านทางโทรศัพท์และทำการจัดส่งสินค้ามาที่บ้าน

4.2 อุปกรณ์ที่ใช้ในการทดลอง

หน่วยประมวลผลกลาง (CPU)	:	Intel Core i7-2670QM @2.20GHz
หน่วยความจำหลัก (Hard Disk)	:	1 TB (1,000 GB)
หน่วยความจำสำรอง (RAM)	:	8 GB
ระบบปฏิบัติการ (OS)	:	Windows 7
ภาษาโปรแกรมที่ใช้	:	PHP 5.3.6 / HTML / XML / XSLT / XPATH
PHP Class Library ที่นำมาใช้	:	PHP Tidy Node

4.3 การวัดประสิทธิภาพในการทดลองของโมเดล

ตัวชี้วัดการทดลองในงานวิจัยนี้ได้เลือกใช้ค่าความแม่นยำ (Precision) ค่าความสามารถในการจดจำ (Recall) และค่าเอฟเมเชอร์ (F-Measure) ในการวัดประสิทธิภาพของโมเดล ซึ่งได้รับความนิยมอย่างมากในงานทางด้านการจำแนกข้อมูล (Classification) โดยการวัดประสิทธิภาพว่าโมเดลไหนดีกว่านั้นจะขึ้นอยู่กับวัตถุประสงค์ของงานในแต่ละด้าน ซึ่งค่าผลลัพธ์ที่ได้ควรสมดุลกันทั้ง Precision และ Recall จะดีที่สุด

4.3.1 ค่าความแม่นยำ (Precision)

ค่าความแม่นยำ หรือ Precision ใช้วัดความแม่นยำในการสกัดข้อมูลของโมเดล โดยวัดจากอัตราส่วนระหว่างเอกสารที่โมเดลสามารถสกัดออกมาได้และตรงกับความต้องการ โดยคิดออกมาเป็นเปอร์เซ็นต์ ดังสมการที่ 4.1

$$Precision = \frac{Correct}{Correct+Incorrect} \quad (4.1)$$

โดยที่ Correct คือโหนดที่ถูกสกัดออกมาได้ถูกต้อง, Incorrect คือ โหนดที่ถูกสกัดออกมาได้แต่ไม่ถูกต้อง

4.3.2 ค่าความสามารถในการจดจำ (Recall)

ค่าความสามารถในการจดจำ หรือ Recall ใช้เพื่อวัดประสิทธิภาพการสกัดข้อมูลของโมเดล โดยวัดอัตราส่วนระหว่างสิ่งที่โมเดลสกัดออกมาได้ถูกต้อง กับเอกสารทั้งหมดที่นำมาทดสอบ ดังสมการที่ 4.2

$$Recall = \frac{Correct}{Relevant} \quad (4.2)$$

โดยที่ Correct คือโหนดที่สกัดออกมาได้และถูกต้อง Relevant คือ โหนดที่ถูกต้องทั้งหมด (รวมทั้งที่สกัดออกมาได้ และสกัดไม่ได้)

4.3.3 ค่าเอฟเมเชอร์ (F-Measure)

ค่าเอฟเมเชอร์ หรือ F-Measure (สามารถเรียกได้อีกหลายชื่อว่า F1, F-Score) ถูกใช้เพื่อวัดประสิทธิภาพระหว่างโมเดลมากกว่า 1 โมเดลเพื่อเปรียบเทียบว่าโมเดลใดมีประสิทธิภาพมากกว่ากัน ดังสมการที่ 4.3

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4.3)$$

โดยที่ Precision คือค่าความแม่นยำที่ได้จากสมการที่ 4.1 และ Recall คือค่าความสามารถในการจดจำข้อมูลที่ได้จากสมการที่ 4.2

4.4 ผลการทดลอง

ตารางที่ 4.2 ผลการทดลองของงานวิจัยนี้ (Subject Detection and Node Density : SDND)

Dataset	Extracted Page	Relevant Node	Correct Node	Incorrect Node	Recall	Precision	F-Measure
Hngcase	150	8250	8250	900	100.00%	90.16%	94.83%
Nadzproject	150	2550	2550	464	100.00%	84.61%	91.66%
Ksscom	150	3900	3768	0	96.62%	100.00%	98.28%
OLX	150	1500	1500	0	100.00%	100.00%	100.00%
ASOS	150	4050	3882	0	95.85%	100.00%	97.88%
JIB Computer	150	7050	6595	949	93.55%	87.42%	90.38%
360TVShopping	147	3969	3969	0	100.00%	100.00%	100.00%
Global Reebok	150	1350	1339	882	99.19%	60.29%	74.99%
Overall	1197	32619	31853	3195	98.15%	90.31%	93.50%

จากตารางที่ 4.2 Dataset คือชื่อโดเมนของเว็บไซต์ที่นำมาทดลอง Extracted Page คือ จำนวนของหน้าเว็บที่ถูกเลือกมาใช้ในการทดลอง โดยได้จากการสุ่มตามหมวดต่างๆของผลิตภัณฑ์ Relevant Node คือ จำนวนของโหนดที่ต้องการ โดยวิธีการได้มาถูกกล่าวไว้แล้วในย่อหน้าแรกในหัวข้อ 4.3 เรื่องการวัดประสิทธิภาพในการทดลอง Correct Node คือ โหนดที่โมเดลสกัดออกมาได้และถูกต้องตรงกับ Relevant Node ส่วน Incorrect Node คือ โหนดที่สกัดออกมาได้ แต่ไม่ตรงกับ Relevant Node

จากตารางที่ 4.2 จะเห็นว่า ค่า Recall เกิน 90 เปอร์เซ็นต์ในทุกเว็บไซต์ เนื่องจากโหนดที่ถูกสกัดได้นั้นจะมีโหนดที่ต้องการรวมอยู่ด้วยเสมอ ตามข้อสังเกตที่ได้กล่าวไว้แล้วว่า โหนดเนื้อหา จะอยู่ใกล้กับโหนดหัวข้อ ส่วนค่า Precision ของเว็บไซต์ Global.Reebok ที่ลดลงถึง 60 เปอร์เซ็นต์ นั้นเนื่องจากว่า โมเดลมองเห็นลิ้งค์เป็นสิ่งรบกวน และเมื่อพบสิ่งรบกวนมากโมเดลก็จะหยุดทำงาน และในเว็บไซต์นั้นมีลิ้งค์รวมอยู่ในส่วนของโหนดเนื้อหาด้วย ตามรูปที่ 4.4 ซึ่งก็คือส่วนของ Breadcrumb ซึ่งทางเว็บไซต์ตั้งใจใส่ไว้เพื่อให้ Web Bot ทั่วไปใช้สร้างไซต์แมพได้ง่าย (สามารถสังเกตุได้ที่ด้านซ้ายบนของรูปที่ 4.4 [< Back / Home / Men / Shoes / ...]) ซึ่งส่วนนี้ทางโมเดลถือว่าเป็น Incorrect Node สำหรับเว็บไซต์ที่ได้ผลลัพธ์ F-Measure ถึง 100 เปอร์เซ็นต์นั้น (OLX, 360TVShopping) เนื่องมาจากในหน้าเว็บไซต์เหล่านั้นมีลิ้งค์อยู่น้อย และมีการแยกแยะส่วนของรายละเอียดผลิตภัณฑ์ออกมาอย่างชัดเจน ดังรูปที่ 4.5 และ รูปที่ 4.8 (สำหรับผลการทดลองในรูปแบบ XML ไฟล์ทั้งหมด สามารถดูได้ที่ ภาคผนวก)

4.5 เปรียบเทียบผลการทดลอง

ในการเปรียบเทียบผลการทดลองได้นำเอางานวิจัยของ Sun, F. Song, D. and Liao, L. ในเรื่อง DOM Based Content Extraction via Text Density (CECTD-DS) [14] เข้ามาเปรียบเทียบด้วย เนื่องจากการสกัดข้อมูลเป็นแบบอัตโนมัติ ในการสกัดข้อมูลซึ่งเหมาะสมกับการนำไปใช้งานจริงเพื่อลดภาระของผู้ใช้งานปลายทางให้มากที่สุด แต่วัตถุประสงค์ของ CECTD-DS จะเน้นไปที่การสกัดข้อมูลจากหน้าเว็บประเภทบทความมากกว่า จึงได้รับผลลัพธ์ที่ค่อนข้างต่ำ และมีอีกโมเดลหนึ่งที่มีวัตถุประสงค์ใกล้เคียงกับงานวิจัยชิ้นนี้ คืองานวิจัยของ Zheng, X. Gu, Y. and Li, Y. ในเรื่อง Data Extraction from Web Pages Based on Structural Semantic Entropy (DE-SSE) [18] ซึ่งมีเป้าหมายในการสกัดข้อมูลจากหน้าเว็บไซต์ประเภทซื้อขายสินค้าออนไลน์ แต่ DE-SSE ไม่สามารถทำงานเป็นอัตโนมัติได้เนื่องจากว่ายังต้องการคีย์เวิร์ดในการทำงานจากผู้ใช้งานอยู่ และเป็นปัญหาที่สำคัญ เพราะหากเปลี่ยนเว็บไซต์ หรือเว็บไซต์ที่ใช้คนละภาษา โมเดลนี้ก็จะเกิดปัญหาทันที

ตารางที่ 4.3 ผลการเปรียบเทียบระหว่าง SDND (งานวิจัยชิ้นนี้) กับ CECTD-DS

Dataset	SDND Tag Feature (F-Measure)	SDND CSS Feature (F-Measure)	SDND Average (F-Measure)	CECTD-DS (F-Measure)
Hngcase	94.83%	94.83%	94.83%	93.04%
Nadzproject	91.66%	91.66%	91.66%	81.63%
Ksscom	98.28%	0.00%	0.00%	72.11%
OLX	100.00%	100.00%	100.00%	42.28%
ASOS	97.88%	97.88%	97.88%	81.41%
JIB Computer	90.38%	90.38%	90.38%	0.00%
360TVShopping	100.00%	60.00%	100.00%	62.07%
Global Reebok	75.00%	75.00%	0.00%	75.00%
Overall	93.50%	76.22%	71.84%	63.44%

จากตารางที่ 4.3 เป็นการเปรียบเทียบผลการทดลองโดยใช้ค่า F-Measure ซึ่งเป็นค่าเฉลี่ยระหว่างความแม่นยำและความสามารถในการจดจำของโมเดลดังที่ได้กล่าวไปแล้วในวิธีการวัดผลการทดลอง ในส่วนของคอลัมน์ SDND Tag Feature และ SDND CSS Feature และ SDND Average คือการทดลองของงานวิจัยนี้ที่ใช้การกำหนดค่าคงที่ต่างกัน โดย Tag Feature คือการกำหนดค่าพื้นฐานดังที่กล่าวไปแล้ว (Default) โดยเราจะกำหนดค่าให้กับชื่อของแท็กมีน้ำหนักมากที่สุด ซึ่งจะกำหนดค่าตามตารางที่ 4.4 ดังนี้ α = 0.5, β = 0.1, γ = 0.1, δ = 0.3, κ = 0.15, λ = 0.15, และ μ = 0.70 คอลัมน์ SDND CSS Feature คือการกำหนดค่าโดยสลับค่าให้คุณสมบัติ CSS มีน้ำหนักมากที่สุด (ขนาดตัวอักษร ความหนาตัวอักษร การจัดเรียงแท็ก) โดยจะกำหนดค่าดังนี้ α = 0.3, β = 0.1, γ = 0.1, δ = 0.5, κ = 0.1, λ = 0.1, และ μ = 0.8 คอลัมน์ SDND Average คือการตั้งค่าคงที่โดยเฉลี่ยค่าเท่ากันในทุกตัวแปรที่อยู่ในระดับเดียวกัน โดยจะตั้งค่าดังนี้ α = 0.25, β = 0.25, γ = 0.25, δ = 0.25, κ = 0.33, λ = 0.33, และ μ = 0.34 และคอลัมน์สุดท้ายเป็นผลการทดลองของโมเดล CECTD-DS ที่ถูกนำมาเปรียบเทียบ จากตารางที่ 4.3 ซึ่งแสดงให้เห็นว่า CECTD-DS ไม่สามารถสกัดข้อมูลรายละเอียดผลิตภัณฑ์จากเว็บไซต์ JIB Computer ได้เลย เนื่องมาจาก CECTD-DS นั้นขึ้นกับความหนาแน่นของตัวอักษรในแต่ละโหนด และในหน้าเว็บไซต์รายละเอียดผลิตภัณฑ์ของ JIB Computer นั้นมีส่วนท้ายของแต่ละเว็บเพจที่บรรจุตัวอักษรอยู่มาก ดังรูปที่ 4.7 ด้านล่าง ทำให้โมเดล CECTD-DS ตรวจโหนดเนื้อหาผิดพลาดซึ่งเป็นส่วนของ About Us และในส่วนท้ายนั้นจะมีอยู่ในทุกๆหน้าเว็บของเว็บไซต์ JIB Computer จึงทำให้ทุกหน้าเว็บนั้นเกิดความผิดพลาดแบบเดียวกันหมด จึงได้รับผลลัพธ์เป็น 0 เปอร์เซ็นต์ในการสกัดข้อมูล และในส่วนที่ได้เปอร์เซ็นต์ต่ำกว่า 50 เปอร์เซ็น (OLX) เนื่องจาก CECTD-DS ตรวจจับโหนดที่ยังไม่ครอบคลุมเนื้อหาทั้งหมด ดังในรูปที่

4.5 นั่นคือ สามารถจับเนื้อหารายละเอียดของผลิตภัณฑ์ได้ แต่ผิดพลาดในส่วนของรูปภาพและช่องทางการติดต่อกับผู้ประกาศขายสินค้าไป

ตารางที่ 4.4 การกำหนดค่าคงที่เพื่อทดลองในงานวิจัยนี้

ตัวแปรค่าคงที่	ค่าน้ำหนัก		
	Tag Featured	CSS Featured	Average
Tag Name (α)	**0.5**	0.3	**0.25**
Title Similarity (β)	0.1	0.1	**0.25**
Keyword Similarity (γ)	0.1	0.1	**0.25**
CSS Properties (δ)	0.3	**0.5**	**0.25**
|- Display (κ)	0.15	**0.1**	0.33
|- Font Weight (λ)	0.15	**0.1**	0.33
|- Font Size (μ)	0.70	**0.8**	0.34

สำหรับข้อจำกัดของงานวิจัยนี้ คือ การสกัดข้อมูลจากหน้าเว็บประเภทซื้อขายสินค้าออนไลน์ที่มีส่วนของรายละเอียดของรายการมากกว่าหนึ่งส่วน เช่น eBay และ Amazon โมเดลนี้จะถูกขัดจังหวะโดยส่วนการแนะนำผลิตภัณฑ์ (Recommend Product) ที่คั่นอยู่ระหว่าง รายละเอียดหลัก (Main Detail) และ รายละเอียดรอง (Sub Detail) ของรายการนั้นๆ จึงทำให้สามารถสกัดได้เพียงแค่ส่วนรายละเอียดหลักเท่านั้น ดังรูปที่ 4.9 ซึ่งในส่วนของกรอบใหญ่คือส่วนที่โมเดลสกัดได้ โดยแบ่งเป็นส่วนบนคือรายละเอียดหลัก และ ส่วนล่างคือรายละเอียดรอง ในส่วนของกรอบเล็กคือส่วนของการแนะนำผลิตภัณฑ์ และอีกข้อจำกัดหนึ่งของโมเดลนี้ก็คือ ไม่สามารถเก็บข้อมูลได้เต็ม 100 เปอร์เซ็นต์เนื่องจากทางผู้พัฒนาเว็บไซต์ตั้งใจสร้างบางส่วนของรายละเอียดให้สะดุดตาผู้เข้าชม เช่น ราคาของผลิตภัณฑ์ ดังนั้นในบางที Tag ที่เก็บราคาจะถูกแยกออกมาอย่างเห็นได้ชัดดังรูปที่ 4.7 หน้าเว็บของ JIB Computer ที่แสดงให้เห็นรายละเอียดอยู่ในกรอบเล็กทางขวา แต่โมเดลนี้สกัดได้ทั้งหมดซึ่งก็รวมถึงส่วนของการแนะนำผลิตภัณฑ์ด้วย

รูปที่ 4.9 ตัวอย่างข้อจำกัดของงานวิจัยชิ้นนี้

บทที่ 5
สรุปและข้อเสนอแนะ

5.1 สรุป

งานวิจัยนี้ได้เสนอวิธีการใหม่ในการสกัดข้อมูลรายละเอียดผลิตภัณฑ์จากหน้าเว็บประเภทซื้อขายสินค้าออนไลน์โดยใช้วิธีการตรวจจับโหนดหัวข้อและความหนาแน่นของโหนด (Subject Detection and Node Density : SDND) ซึ่งงานวิจัยนี้สามารถประยุกต์ใช้กับการทำเหมืองข้อมูลเว็บไซต์ (Web Mining) ในส่วนของการจัดเตรียมข้อมูล (Pre-Processing) ซึ่งเป็นประโยชน์กับงานประยุกต์ในหลายรูปแบบ เช่น การวิเคราะห์ผลิตภัณฑ์เชิงธุรกิจ ระบบสนับสนุนการตัดสินใจ ระบบแนะนำผลิตภัณฑ์ เป็นต้น โดยการทำงานของโมเดลนี้จะเริ่มต้นที่การตรวจจับหัวข้อก่อน โดยวิเคราะห์จากการคำนวณหาค่าน้ำหนักมวลรวมของทุกแท็กที่มีความเป็นไปได้ว่าจะเป็น โหนดหัวข้อ หลังจากนั้นจึงทำการตั้งเป็นโหนดปัจจุบัน และทำการคำนวณหาค่าอัตราความหนาแน่นของโหนดพร้อมกับ เดินย้อนกลับไปยังโหนดแม่ และทำซ้ำไปเรื่อยๆจนกระทั่งค่าอัตราการเพิ่มขึ้นของโหนดปัจจุบันเพิ่มขึ้นแบบผิดปกติ จึงหยุดทำงานและเลือกโหนดปัจจุบันเป็นโหนดรายละเอียดผลิตภัณฑ์ทันที โดยที่งานวิจัยนี้ได้เลือกแสดงผลลัพธ์ในรูปแบบของไฟล์ XML เพื่อให้ดูง่ายและสามารถเรียกค้นได้ง่าย ข้อได้เปรียบหรือข้อดีของงานวิจัยนี้ก็คือ เป็นวิธีการแบบอัตโนมัติ โดยที่โมเดลไม่ต้องการตัวแปรใดๆ ช่วยในการคำนวณ ต้องการเพียงแค่ URL ของหน้าเว็บที่ต้องการสกัดเพียงเท่านั้น ซึ่งจะมีความเหมาะสมกับผู้ใช้งานปลายทาง เนื่องจากผู้ใช้งานปลายทางไม่มีความเชี่ยวชาญใดๆ จึงง่ายต่อการนำไปใช้ ส่วนข้อจำกัดของงานวิจัยนี้คือ เป็นโมเดลที่ขึ้นกับโหนดหัวข้อ ดังนั้นหากมีความผิดพลาดในการตรวจจับโหนดหัวข้อจะทำให้ผลลัพธ์ผิดเพี้ยนไปมาก และยังติดปัญหากับเว็บไซต์บางประเภทที่มีส่วนของการแนะนำผลิตภัณฑ์รวมอยู่ในส่วนของรายละเอียดผลิตภัณฑ์ และอีกปัญหาหนึ่งคือบางเว็บไซต์ทำการแยกโหนดข้อมูลบางโหนดออกไปจากกลุ่มของโหนดรายละเอียด จึงทำให้โมเดลนี้สกัดข้อมูลได้ไม่ครบถ้วน

5.2 ข้อเสนอแนะ

สำหรับการนำงานวิจัยนี้ไปใช้งานจริง สามารถทำได้โดยการนำอัลกอริธึมไปสร้างเป็นโปรแกรมตามภาษาที่ถนัด (สำหรับการทดลองในงานวิจัยนี้ได้ใช้ภาษา PHP และ Tidy Node Library) โดยการนำไปใช้เพียงแค่กำหนดค่าชื่อแท็กที่ต้องการ (ค่าพื้นฐานคือ <H1>, <H2>, <H3>, <DIV> และ <SPAN>) โดยหากไม่ต้องการตรวจจับหัวข้อของโหนดเนื้อหาก็สามารถเปลี่ยนชื่อแท็กเหล่านี้ได้ และการกำหนดค่าอีกอย่างหนึ่งคือค่าคงที่ที่ใช้ถ่วงน้ำหนักตามความสำคัญของแต่ละคุณลักษณะที่ใช้ โดยหากผู้นำไปใช้งานต้องการวัตถุประสงค์เดียวกัน คือ สกัดโหนดเนื้อหาจากหน้า

เว็บประเภทซื้อขายสินค้าออนไลน์ในหน้าเว็บที่มีข้อมูลอยู่หนาแน่น ทางผู้ใช้งานสามารถใช้ค่าพื้นฐานตามที่งานวิจัยนี้กำหนดไว้และนำไปสร้างเป็นโปรแกรมเพื่อใช้งานได้ทันที

ส่วนงานวิจัยในอนาคตนั้น จะนำไปทดลองกับเว็บไซต์ประเภทอื่น เช่น เว็บประเภทบทความหรือ กระดานสนทนา เพื่อทดลองความยืดหยุ่นของโมเดล และแก้ไขข้อจำกัดในโมเดลนี้ อีกส่วนหนึ่งคือปรับปรุงแนวความคิดเพื่อนำไปใช้กับงานด้านการสกัดความคิดเห็นบนเว็บไซต์ต่างๆ เพื่อใช้ในการทำนายความคิดเห็น (Opinion Mining) เป็นต้น

บรรณานุกรม

[1] Bhardwaj, A. and Mangat, V. 2014. **A Novel Approach for Content Extraction from Web Pages.** Engineering and Computational Sciences (RAECS). pp. 1-4.

[2] Brauer, F. Robert, R. Adrian, M. and Wojciech, M.B. 2011. **Enabling Information Extraction by Inference of Regular Expressions from Sample Entities.** ACM International Conference on Information and Knowledge Management (CIKM). pp. 1285-1294.

[3] Cai, D. Yu, S. Wen, J. and Ma, W. 2003. **VIPS: a Vision-based Page Segmentation Algorithm.** Microsoft Research. pp. 1-29.

[4] Chainapaporn, P. Netisopakul, P. 2012. **Thai Herb Information Extraction from Multiple Websites.** Interntional Conference on Knowledge and Smart Technology (KST). pp. 16-23.

[5] Chang, C. Hsu, C. and Lui, S. 2001. **IEPAD: Information Extraction Based on Pattern Discovery.** Proceedings of the 10th international conference on World Wide Web. pp. 681–688.

[6] Chang, C. Kayed, M. Girgis, M. and Shaalan, K. 2006. **A Survay of Web Information Extraction System.** IEEE Transactions on Knowledge and Data Engineering. pp. 1411-1428.

[7] Dias, S. and Gadge, J. **Identifying Informative Web Content Blocks using Web Page Segmentation.** International Journal of Applied Information Systems (IJAIS). pp. 37-41.

[8] Fumarola, F. Weninger, T. and Barber, R. 2011. **Extracting General Lists from Web Documents: A Hybrid Approach (HyLiEn).** International Conference on Industrial Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE). pp. 285-294.

[9] Jellouli, I. and Mohajir, M. 2009. **Towards automatic semantic annotation of data rich Web pages.** Research Challenges in Information Science. pp. 139-142.

[10] Laender, A.H.F. Ribeiro-Neto, B.A. da Silva, A.S. Teixeira, J.S. 2002. **A Brief Survey of Web Data Extraction Tools.** Special Interest Group on Management of Data (SIGMOD). pp. 84-93.

[11] Pusdekar, S.J. and Chhaware, S.P. 2014. **Using Visual Clues Concept for Extracting Main Data from Deep Web Pages.** International Conference on Electronic Systems, Signal Processing and Computing Technologies. pp. 190-193.

[12] Schreng, M. 2012. **Webbots Spiders and Screen Scrapers : A Guild to Developing Internet Agents with PHP/CURL.** 2nd ed. San Francisco : No Starch Press.

[13] Sleiman, H. and Conchuelo, R. 2013. **A Survey on Region Extractors from Web Documents.** IEEE Transactions on Knowledge and Data Engineering. pp. 1960-1981.

[14] Sun, F. Song, D. and Liao, L. 2011. **DOM Based Content Extraction via Text Density.** Special Interest Group on Information Retrieval. pp. 245-254.

[15] Thamviset, W. and Wongthanavasu, S. 2012. **Structured Web Information Extraction Using Repetitive Subject Pattern.** Electrical Engineering/Electronics Computer Telecommunications and Information Technology (ECTICON). pp. 1-4.

[16] Thamviset, W. and Wongthanavasu, S. 2013. **Information Extraction for Deep Web Using Repetitive Subject Pattern.** Machine Learning and Intelligent System (MLIS). pp. 1109-1139.

[17] Wang, J. and Loshovsky, F. 2002. **Data-rich Section Extraction from HTML pages.** Web Information Systems Engineering (WISE). pp. 1-10.

[18] Zheng, X. Gu, Y. and Li, Y. 2012. **Data Extraction from Web Pages Based on Structural-Semantic Entropy.** International World Wide Web Conference Committee (W3C2). pp. 93-102.

ภาคผนวก ก

ผลการทดลองในรูป XML

ตัวอย่างผลลัพธ์จากแต่ละเว็บไซต์

```
<store name="hngcase" crawl_date="2014-12-13">
    <item>
        <subject>
            iPhone 6 Nillkin Amazing CP+ Full Screen Cover Glass Screen Protector White
        </subject>
        <url>
            http://www.hngcase.com/product/2925/iphone-6-nillkin-amazing-cp-full-
            screen-cover-glass-screen-protector-white
        </url>
        <detail>
            <d0>หน้าแรก</d0>
            <d1>&gt;</d1>
            <d2>สินค้าทั้งหมด</d2>
            <d3>Case iPhone</d3>
            <d4>Case iPhone 6</d4>
            <d5>ฟิล์มกันรอย</d5>
            <d6>iPhone 6 Nillkin Amazing CP+ F...</d6>
            <d7>หมวดหมู่</d7>
            <d8>ราคาปกติ</d8>
            <d9>690.00</d9>
            <d10>บาท</d10>
            <d11>ลดเหลือ</d11>
            <d12>550.00 บาท</d12>
            <d13>สถานะสินค้า</d13>
            <d14>พร้อมส่ง</d14>
            <d15>ประเภท</d15>
            <d16>สินค้าใหม่</d16>
            <d17>อัพเดทล่าสุด</d17>
            <d18>11 ธ.ค. 2557</d18>
            <d19>ความพึงพอใจ</d19>
            <d20>ยังไม่มีความคิดเห็น</d20>
            <d21>จำนวน</d21>
            <d22>ชิ้น</d22>
            <d23>หยิบลงตะกร้า</d23>
            <d24>Tweet</d24>
            <d25>รายละเอียดสินค้า</d25>
            <d26>วิจารณ์ (0)</d26>
            <d27>ถาม - ตอบ (0)</d27>
            <d28>วิธีการชำระเงิน</d28>
            <d29>เงื่อนไขอื่นๆ</d29>
```

<d30>
ซื้อสินค้า 2 รายการขึ้นไปจัดส่งฟรีEMS ฟิล์มกระจกนิรภัยกันจอแตก เต็มหน้าจอ Transparent ความหนา : 0.3 mm. ประเภท : ฟิล์มกระจกนิรภัย รุ่น :Amazing H+Anti-Explosion Glass วัสดุ :Japan imported AGC glass material อุปกรณ์ในกล่อง : ฟิล์มเต็มหน้าจอ + ฟิล์มเลนส์กล้อง + อุปกรณ์ติดฟิล์มพร้อมผ้าเช็ดหน้าจอ รายละเอียดสินค้า ฟิล์มกระจกนิรภัย iPhone 6 แบบเต็มหน้าจอใช้วัสดุนำเข้าจากญี่ปุ่น (AGC glass) ติดตั้งได้ง่าย มีให้เลือก สีขาว และ ดำ ตามบอดี้เครื่องให้การปกป้องที่ดีเยี่ยม ช่วยให้สีสันหน้าจอสดใสตามจริงยิ่งขึ้นลดแสงสะท้อนได้ขณะใช้งานกลางแจ้ง ด้วยความแข็งระดับ 9Hทำให้ทนทางต่อรอยขีดข่วน ลูกกุญแจ มีดกรีด ทั้งยังช่วยลดรอยนิ้วมืออีกด้วยหนาเพียง 0.3 mm.ทำความสะอาดได้ง่าย 2.5Dตัดขอบฟิล์มให้โค้งมนยิ่งขึ้นทำให้เนื้อฟิล์มบางแนบกับหน้าจอได้สนิทเสมือนไม่ได้ใส่ฟิล์ม และไม่คมทำให้ปลอดภัยไม่บาดมือในขณะติดฟิล์ม
</d30>
<d31>คลิปแนะนำสินค้า</d31>
<d32>วิธีติดฟิล์มกระจกนิรภัยNillkin Amazing H+</d32>
<d33>สินค้านี้ยังไม่มีคนวิจารณ์</d33>
<d34>ชื่อ</d34>
<d35>อีเมล</d35>
<d36>คำถาม</d36>
<d37>รายละเอียด</d37>
<d38>ถาม</d38>
<d39>สินค้านี้ยังไม่มีคนถามคำถาม</d39>
<d40>
สามารถโอนเงินได้ทันทีโดยไม่ต้องรอคอนเฟิร์มจากทางร้าน เนื่องจากเป็นสต็อคสินค้า ออนไลน์แบบ Real Time สินค้าที่ไม่ได้ขึ้นว่าหมดคือมีพร้อมส่งค่ะ
</d40>
<d41>
ข้อแนะนำ ควรโอนติดเศษทศนิยม เช่น ยอดชำระ 300 บาท โอน 300.02 บาท เพื่อความสะดวกและรวดเร็วในการตรวจสอบ
</d41>
<d42>
ร้านHNGCASEตัดยอดแจ้งชำระเงินเวลา 9.00 น.วันจันทร์-ศุกร์แจ้งชำระเงินภายในเวลาดังกล่าวจะจัดส่งในรอบวันนั้น หากแจ้งชำระเงินหลังเวลาดังกล่าวจะจัดส่งในรอบถัดไปค่ะ
</d42>
<d43>ชำระเงินผ่านทางธนาคาร</d43>
<d44>ธนาคาร</d44>
<d45>สาขา</d45>
<d46>เลขที่บัญชี</d46>
<d47>ประเภทบัญชี</d47>
<d48>ชื่อบัญชี</d48>
<d49>ไทยพาณิชย์</d49>

```
                <d50>เซ็นทรัลพลาซ่าลาดพร้าว</d50>
                <d51>206-232618-5</d51>
                <d52>ออมทรัพย์</d52>
                <d53>วัลย์ลิกา หารต๊ะ</d53>
                <d54>กสิกรไทย</d54>
                <d55>730-2-84291-9</d55>
                <d56>กรุงเทพ</d56>
                <d57>087-7-11792-9</d57>
                <d58>กรุงไทย</d58>
                <d59>690-0-22392-3</d59>
                <d60>ทหารไทย</d60>
                <d61>233-2-25578-1</d61>
                <d62>ไปหน้าแจ้งชำระเงิน</d62>
                <d63>
                        เพื่อความเข้าใจตรงกันระหว่างผู้ซื้อและผู้ขายและมิตรภาพที่ดีต่อกัน  กรุณา
                    ทำความเข้าใจเงื่อนไข
                </d63>
            </detail>
        </item>
        .
        .
        .
        <time_second>286.51538800</time_second>
</store>
```

รูปที่ ก.1 ตัวอย่างผลการทดลองจากเว็บไซต์ HNGCase

```
<store name="nadzproject" crawl_date="2014-12-13">
        <item>
            <subject>PS4: Travel Case (Third Party)</subject>
            <url>
                http://www.nadzproject.com/product.detail_1041995_th_5931559
            </url>
            <detail>
                <d0>
                    ----------------------------------------------------------------------------------------------------
                    ----------
                </d0>
                <d1>**กรุณารอตอบกลับจากเมล์</d1>
                <d2>Order@nadzproject.com</d2>
                <d3>ฉบับที่ 2</d3>
                <d4>ก่อนทำการโอนเงินเข้ามาทุกครั้งนะครับ</d4>
                <d5>**</d5>
```

<d6>รายละเอียดสินค้า/บริการ</d6>
<d7>
สินค้า/บริการ >> Playstation 4 >> Accessories >> Others >> PS4: Travel Case (Third Party)
</d7>
<d8/>
<d9>รหัสสินค้า:</d9>
<d10>005587</d10>
<d11>ราคา:</d11>
<d12>790.00</d12>
<d13>บาท</d13>
<d14>รายละเอียด:</d14>
<d15>
กระเป๋าใส่เครื่อง PS4 เหมาะสำหรับพกไปเล่นตามที่ต่างๆครับ
</d15>
<d16>นโยบายของเว็บไซต์ (Merchant Policy)</d16>
<d17>วิธีการชำระเงิน (Payment Method) อ่านทั้งหมด</d17>
<d18>โอนเงินผ่านธนาคาร</d18>
<d19>การส่งสินค้า (Shipping information)</d19>
<d20>วิธีการสั่งซื้อสินค้าผ่านไปรษณีย์</d20>
<d21>
1. สั่งซื้อสินค้าผ่านหน้าเว็บไซต์www.nadzproject.com โดย add สินค้าเข้าที่ตะกร้ารถเข็น หลังกดสั่งซื้อ (กรุณาออีเมล์ฉบับที่ 2) ทางร้านจะเช็คสินค้าและแจ้งรายละเอียดให้ทราบทางอีมเล์ order@nazproject.com ค่อยทำการชำระค่าสินค้าเข้ามาครับ 2. อีเมล์แจ้งสั่งซื้อโดยตรง เขียนสินค้าที่ต้องการเข้ามาแจ้งที่ order@nadzproject.com 3.สั่งซื้อทางไลน์ Line ID : order_nz (จันทร์ - ศุกร์ 10.00-19.00 น. / เสาร์ 12.30-18.00 น) 4.สั่งซื้อทางเพจ Facebook inbox : www.facebook.com/nadzonlinestore 5.ติดต่อสอบถาม 02-2353036 ต่อ 106 และ 107 มือถือ 098-2724403
</d21>
<d22>รายละเอียดบริษัท (Company Detail)</d22>
<d23>
แน๊สโปรเจ็ค จำหน่ายเครื่องเกมส์ทุกคอนโซล อุปกรณ์เกมส์ทุกประเภท จัดส่งสินค้าทั่วประเทศรวดเร็วผ่านไปรษณีย์ไทย 1.ดิจิตอลเกทเวย์ สยามแสควร์ ชั้น5 [SHOP] 2.เซ็นทรัลลาดพร้าว ชั้นG [SHOP] 3.อาคาร ITFสีลมพาเลส ชั้น14 [LOGISTIC] ศูนย์รวมเครื่องเล่นเกมส์และอุปกรณ์เกมส์คอนโซลทุกชนิด เราจำหน่ายทั้งราคาปลีกและราคาส่ง นอกจากนั้นเราย... อ่านทั้งหมด
</d23>
<d24>นโยบายการคืนสินค้า (Return Policy)</d24>
<d25>

```
                สั่งซื้อผ่านไปรษณีย์ สินค้ามีประกัน 7 วันเปลี่ยน นับจากวันที่ลูกค้าได้รับ สอบถาม
                ข้อมูลการจัดส่งติดต่อ 02-235-3036# 106 , 098-2724403 FAX: 02-235-3036
                ต่อ 8 ได้ในเวลาทำการ
            </d25>
            <d26>เวลาให้บริการ (Business Hour)</d26>
            <d27>
                ลูกค้าทุกท่านสามารถสั่งซื้อผ่านเว็บไซต์ได้ตลอด 24 ชั่วโมง โดยจะได้รับการตอบรับ
                กลับไม่เกิน 24Hr ภายในวันเวลาทำการ สามารถสอบถามข้อมูลบริการ ปัญหา
                สั่งซื้อสินค้า และ แจ้งการโอนเงินได้ที่ order@nadzproject.com เวลาเปิดทำการ
                แยกตามสาขาร้าน Nadzproject สาขาสยาม จันทร์ - อาทิตย์ : 11:00น - 20:00น
                02-623-4158 , 086-456-8632 สาขาลาดพร้าว จันทร์ - อาทิตย์ : 11:00น -
                21:30น 02-937-0936 , 087-678-6663 Nadz Online Store (ฝ่ายจัดส่ง)จันทร์ -
                ศุกร์ : 10:00น - 19:00น / เสาร์ : 13:00น - 18:00 น / อาทิตย์ : ปิดจัดส่ง 02-
                235-3036 # 106 , 098-2724403 NADZCARE(ศูนย์ซ่อม) จันทร์ - อาทิตย์ :
                10:30น - 19:30น 02-623-4159 , 086-366-3539 Customer Service (แจ้ง
                พนักงานบริการไม่สุภาพ) จันทร์ - ศุกร์ : 10:00น - 19:00น / เสาร์ : 13:00น -
                18:00 น / อาทิตย์ : ปิดทำการ 02-235-3036 #11 E-Mail :
                customerservice@nadzproject.com
            </d27>
            <d28>การรักษาข้อมูล (Privacy Policy)</d28>
            <d29>
                ทางเว็บไซต์ยึดถือปฏิบัติตามนโยบายรักษาความลับส่วนบุคคลอย่างเคร่งครัดในการ
                รับข้อมูลส่วนตัวของท่าน ผ่านทางเว็บไซต์และทางเว็บไซต์จะไม่ส่งข้อมูลส่วนตัวของ
                ท่านไปยังบุคคลอื่นโดยไม่ผ่านความยินยอม จากท่านโดยเด็ดขาด
            </d29>
        </detail>
    </item>
    .
    .
    .
    <time_second>214.81128700</time_second>
</store>
```

รูปที่ ก.2 ตัวอย่างผลการทดลองจากเว็บไซต์ Nadzproject

```
<store name="ksscom" crawl_date="2014-12-13 15:20:24">
    <item>
        <subject>
            Commy Data cable Flat 8 PIN Lightning รุ่น DC219 สีเทา
        </subject>
        <url>
            http://ksscom.co.th/index.php?component=catalog&call=pr
            oduct_detail&pro_id=2490
```

```
            </url>
            <detail>
                <d0>ปลีก : 190 บาท</d0>
                <d1>ส่ง : เฉพาะลูกค้าประจำ</d1>
                <d2>ดีเลอร์ : เฉพาะดีเลอร์</d2>
                <d3>รหัสสินค้า :</d3>
                <d4>8853201072695</d4>
                <d5>หมวดหมู่ :</d5>
                <d6>Cable Charger</d6>
                <d7>ยี่ห้อ :</d7>
                <d8>COMMY</d8>
                <d9>เริ่มจำหน่าย :</d9>
                <d10>03-12-2014</d10>
                <d11>มีผู้เข้าชมแล้ว :</d11>
                <d12>6 ครั้ง</d12>
                <d13>หยิบใส่ตระกร้า</d13>
                <d14>รายละเอียดเพิ่มเติม</d14>
                <d15>- ความยาวสาย 100 ซม.</d15>
                <d16>- สายแบบแบน หมดปัญหาสายพันกัน</d16>
                <d17>- มีความทนทาน และสีสันสวยงาม</d17>
                <d18>- รองรับการชาร์จ และซิงค์ข้อมูล</d18>
                <d19>- รับประกัน 6 เดือน</d19>
                <d20>- รองรับ ios 8, Iphone6 และ Iphone6 Plus</d20>
                <d21>รูปภาพขนาดใหญ่</d21>
                <d22>แสดงความคิดเห็น</d22>
            </detail>
        </item>
        .
        .
        .
        <time_second>1,824.83137500</time_second>
</store>
```

รูปที่ ก.3 ตัวอย่างผลการทดลองจากเว็บไซต์ KSSCom

```
<store name="olx" crawl_date="2014-12-14 00:52:46">
        <item>
            <subject>Sumsung galaxy grand</subject>
            <url>http://www.olx.co.th/product-102307649</url>
            <detail>
                <d0>ยี่ห้อ : Samsung</d0>
                <d1>อุปกรณ์พร้อมกล่องสายชาท หูฟัง หน้าจอติดฟิลม์กันรอยตลอด</d1>
                <d2>4,500.-</d2>
```

```
                <d3>
                    อ.เมืองเชียงใหม่ จ.เชียงใหม่ วันนี้ 00:40:23  น. กดดูเบอร์โทรผู้ขาย 0802xxxx
                    0802497429
                </d3>
                <d4>
                    กรุณาอย่าโอนเงินไม่ว่าในกรณีใดๆ หากผู้ขายต้องการให้โอนเงินก่อนหรือมัดจำ
                    ล่วงหน้า โปรดแจ้งฝ่ายบริการลูกค้า โทร. 02-833-3187  หรือ cs@olx.co.th
                    ตรวจสอบบัญชีรายชื่อคนโกง ที่นี่!
                </d4>
                <d5>กดดู Line ID</d5>
                <d6>Yuisalola</d6>
                <d7>สมาชิก 2409599</d7>
                <d8>ขาย มือสอง</d8>
            </detail>
        </item>
        .
        .
        .
        <time_second>277.40086700</time_second>
</store>
```

รูปที่ ก.4 ตัวอย่างผลการทดลองจากเว็บไซต์ OLX

```
<store name="asos" crawl_date="2014-12-14 09:48:01">
        <item>
            <subject>ASOS Chunky Hand Knit Funnel Snood</subject>
            <url>
                http://www.asos.com/ASOS/ASOS-Chunky-Hand-Knit-Funnel-
                Snood/Prod/pgeproduct.aspx?iid=4810274&amp;cid=4174&amp;sh=0&amp;pg
                e=0&amp;pgesize=36&amp;sort=-
                1&amp;clr=Khaki&amp;totalstyles=1159&amp;gridsize=3
            </url>
            <detail>
                <d0>£16.00</d0>
                <d1>Free Shipping Worldwide*»</d1>
                <d2>Snood by ASOS Collection</d2>
                <d3>Super soft touch chunky knit</d3>
                <d4>Fluffy feel finish</d4>
                <d5>Funnel styling</d5>
                <d6>Machine wash</d6>
                <d7>100% Acrylic</d7>
                <d8>Total length: 56cm/22&amp;quot;</d8>
                <d9>ABOUT ASOS COLLECTION</d9>
```

<d10>

Directional, exciting and diverse, the ASOS Collection makes and breaks the fashion rules. Scouring the globe for inspiration, our London based Design Team is inspired by fashion's most covetable trends; providing you with a cutting edge wardrobe season upon season.

</d10>

<d11>First select from 1 colours</d11>

<d12>Khaki</d12>

<d13>

Match the fit of this item against your wardrobe favourite

</d13>

<d14>Fit Visualiser</d14>

<d15>Select Size</d15>

<d16>Size Guide</d16>

<d17>Add To Bag</d17>

<d18>NEW - Buy the Look Save For later</d18>

<d19>Info & Care</d19>

<d20>Delivery</d20>

<d21>Returns</d21>

<d22>

ABOUT MEFabric: 100% Acrylic. SIZE & FIT Total length: 56cm/22" Width: 35cm/14" LOOK AFTER MEMachine Wash According To Instructions On Care Label Product Code:

</d22>

<d23>595670</d23>

<d24>

UK -Standard DeliveryDelivery within6 working days (delivery Monday-Saturday)- FREE (spend over £20) otherwise £3 UK -Next Day DeliveryOrder by 10pm Monday-Friday or 5pm Saturday-Sunday for delivery the next day - £5.95 (FREE if you spend over £100) UK - Evening Next DayOrder by midnight for delivery the next evening between 6pm-10pm - £7.95 UK - Nominated DaySelect your delivery day (Monday-Sunday) - £5.95 UK -Click & Collect- Standard DeliveyDelivery within6 working days (delivery Monday-Saturday)- FREE (spend over £20) otherwise £3 UK -Click & Collect- Next Day DeliveryOrder by 6pm Monday-Friday or 5pm Sunday for delivery the next day - £5.95 (FREE if you spend over £100) Click here for more information on UK delivery times

</d24>

<d25>

Click here for more information onInternational delivery times

```
                </d25>
                <d26>
                        Clickhere for information on how to return something to ASOS from
                        within the UK. Clickhere for information on how to return something to
                        ASOS from outside the UK.
                </d26>
                <d27>Tweet</d27>
        </detail>
    </item>
    .
    .
    .
    <time_second>829.55344800</time_second>
</store>
```

รูปที่ ก.5 ตัวอย่างผลการทดลองจากเว็บไซต์ ASOS

```
<store name="jib" crawl_date="2014-12-14 12:36:37">
    <item>
        <subject>
            TABLET SAMSUNG GALAXY TAB 3 7.0 LITE 3G BLACK
        </subject>
        <url>
            http://www.jib.co.th/web/index.php/product/readProduct/13198/43/TABLET-
            SAMSUNG-GALAXY-TAB-3-7.0-LITE-3G-BLACK
        </url>
        <detail>
            <d0>HOME &amp;gt;</d0>
            <d1>
                โทรศัพท์มือถือและแท็บเล็ต &amp;gt; &amp;gt; SAMSUNG &amp;gt; Item#: 8806086044431
            </d1>
            <d2>รายละเอียดสินค้า</d2>
            <d3>คุณสมบัติ</d3>
            <d4>
                TABLET SAMSUNG GALAXY TAB 3 7.0 LITE 3G BLACK - 8806086044431
            </d4>
            <d5>OPERATING SYSTEM</d5>
            <d6>Android OS, v4.2 (Jelly Bean)</d6>
            <d7>PROCESSOR (CPU)</d7>
            <d8>Dual-core 1.2 GHz</d8>
            <d9>GRAPHICS CHIP (GPU)</d9>
            <d10>Vivante GC1000</d10>
            <d11>DISPLAY</d11>
```

<d12>TFT capacitive touchscreen, 16M colors</d12>
<d13>INTERNAL MEMORY</d13>
<d14>8 GB.</d14>
<d15>RAM</d15>
<d16>1 GB.</d16>
<d17>ROM</d17>
<d18>EXTERNAL MEMORY</d18>
<d19>microSD, up to 32 GB.</d19>
<d20>SIZE</d20>
<d21>7"</d21>
<d22>WEIGHT</d22>
<d23>322 g</d23>
<d24>CAMERA</d24>
<d25>Front Camara : No Back Camara : 2 MP</d25>
<d26>CONNECTIVITY</d26>
<d27>
2G Network : GSM 850 / 900 / 1800 / 1900 3G Network : HSDPA 900 / 2100
</d27>
<d28>NETWORK</d28>
<d29>
Wi-Fi 802.11 b/g/n, Wi-Fi Direct, Wi-Fi hotspot
</d29>
<d30>BATTERY</d30>
<d31>Non-removable Li-Ion 3600 mAh battery</d31>
<d32>SENSORES</d32>
<d33>Accelerometer</d33>
<d34>OTHER</d34>
<d35>WARRANTY</d35>
<d36>1</d36>
<d37>
TABLET SAMSUNG GALAXY TAB 3 7.0 LITE 3G BLACK-8806086044431
</d37>
<d38>บาง เบา เปี่ยมสมรรถนะ</d38>
<d39>ดีไซน์เพรียวบาง น้ำหนักเบาเป็นพิเศษ</d39>
<d40>
Tab3 Lite ออกแบบเป็นพิเศษเพื่อความสะดวกในการพกพาติดตัวไปทุกที่ เพรียวบางกว่ารุ่นอื่น ๆ ด้วยความหนาเพียง 9.7 มม. มอบความพึงพอใจในการใช้งานอย่างแท้จริง จับกระชับมือ น้ำหนักเบาเป็นพิเศษ คุณจึงรู้สึกถึงความแตกต่างอย่างชัดเจน ขนาดเล็กกะทัดรัด ประหยัดพลังงาน พร้อมสรรพด้วยคุณสมบัติมากมายอย่างที่คุณต้องการ
</d40>

```
            <d41>ประสบการณ์การรับชมที่เหนือชั้น</d41>
            <d42>
                หน้าจอความละเอียดสูงของ Tab3 Lite ช่วยให้คุณเพลิดเพลินกับความบันเทิงได้
                อย่างเต็มอิ่ม ภาพที่สว่างสดใสและข้อความที่คมชัด โดยไม่มีแถบสถานะที่ปิดกั้นการ
                รับชม ช่วยให้คุณเพลิดเพลินกับพื้นที่การรับชมที่กว้างขวางมากขึ้น
            </d42>
            <d43>โปรเซสเซอร์ Dual Core ความเร็ว 1.2GHz</d43>
            <d44>
                โปรเซสเซอร์ที่แข็งแกร่งและรวดเร็วช่วยให้คุณเพลิดเพลินกับวิดีโอ เกม แอพ และ
                การท่องเว็บได้อย่างไร้ขีดจำกัด เพื่อประสบการณ์มัลติมีเดียบนแท็บเล็ตที่เหนือชั้น
                อย่างแท้จริง
            </d44>
            <d45>กล้องอัจฉริยะ พร้อมฟีเจอร์สนุก</d45>
            <d46>
                กล้องด้านหลังประกอบด้วยฟีเจอร์อัจฉริยะที่ช่วยให้คุณสร้างสรรค์ภาพถ่ายที่
                สมบูรณ์แบบได้ทุกครั้ง ไม่ว่าจะเป็นภาพทิวทัศน์ที่งามตระการตา หรือภาพถ่ายหมู่ที่
                สมบูรณ์แบบ พร้อมรอยยิ้มที่ปรากฏบนใบหน้าของทุกคน คุณก็สามารถถ่ายภาพได้
                อย่างมืออาชีพ และแชร์ภาพได้ทันใจ ด้วยฟีเจอร์การแชร์แบบอัจฉริยะที่มอบความ
                สะดวกให้แก่คุณ
            </d46>
            <d47>บริการสนุก ๆ ที่แบ่งปันกันได้</d47>
            <d48>
                หลายแสนแอพรอให้คุณใช้งานอย่างง่ายดายเพียงปลายนิ้วสัมผัส โดยทั้งหมดนี้
                ออกแบบเป็นพิเศษสำหรับ Tab3 Lite และดาวน์โหลดได้อย่างง่ายดายเพื่อความ
                เพลิดเพลินสูงสุด
            </d48>
            <d49>
                และ Samsung Link จะบันทึกรูปภาพและไฟล์ของคุณไปยังคลาวด์สตอเรจ หรือ
                อุปกรณ์ใด ๆ เพื่อเข้าถึง แก้ไข และแชร์ รวมทั้งจัดการเนื้อหาหรือเล่นวิดีโอบนจอทีวี
                ได้อย่างง่ายดายผ่านการเชื่อมต่อระยะไกล Dropbox จัดเก็บไฟล์ของคุณอย่าง
                ปลอดภัยบนบริการคลาวด์บนเว็บ เพื่อให้เข้าถึงได้ทุกที่ทุกเวลา ง่ายดาย ปลอดภัย
                และไว้ใจได้
            </d49>
        </detail>
    </item>
    .
    .
    .
    <time_second>817.16473900</time_second>
</store>
```

รูปที่ ก.6 ตัวอย่างผลการทดลองจากเว็บไซต์ JIBComputer

```
<store name="360tvshopping" crawl_date="2014-12-14 13:21:39">
    <item>
        <subject>
            แพ็คคู่สุดคุ้ม A ลู่วิ่งไฟฟ้า MTH 6.0L + SIX PACK CARE
        </subject>
        <url>
            http://www.360tvshopping.com/แพ็คคู่สุดคุ้ม-a-ลู่วิ่งไฟฟ้า-mth-6-0lsix-pack-car-
            60998.product
        </url>
        <detail>
            <d0>มีผู้ชม 86 ครั้ง</d0>
            <d1>รหัสสินค้า : P2A</d1>
            <d2>ปกติ :</d2>
            <d3>30,400.00</d3>
            <d4>ลดเหลือ : ฿</d4>
            <d5>27,500.00</d5>
            <d6>THB</d6>
            <d7>จำนวน</d7>
            <d8>รายละเอียด</d8>
            <d9>ลู่วิ่งไฟฟ้า MTH SERIES MTH 6.0L (MTH 6.0L)</d9>
            <d10>
                ลู่วิ่งไฟฟ้า MTH SERIES MTH 6.0L (MTH 6.0)สามารถใช้ออกกำลังกายภายใน
                บ้านได้ประหยัดพื้นที่ในการทำงาน มีระบบบอกการคำนวนระยะทางในการวิ่ง การ
                เต้นของหัวใจ และเครอลี่ที่ใช้ ลู่วิ่งสามารถยกพับเก็บได้ง่าย มีระบบล็อครักษาความ
                ปลอดภัยในการเก็บ เครื่องวิ่งการแสดงผลเป็นระบบหน้าจอ LCD ทำงานด้วยระบบ
                มอเตอร์ ในการหมุนสามารถเพิ่มระดับความลาดเอียงด้วยการควบคุมที่แฮนด์จับ มี
                การออกแบบด้วยเทคโนโลยี นวัตกรรมรุ่นใหม่มีปุ่มหมุนเพื่อช่วยเป็นทางลัดในการ
                เพิ่มความเร็วอย่างต่อเนื่อง รองรับการทำงานด้วยเครื่องเล่น MP3 ด้วยระบบ HI-FI
            </d10>
            <d11>รับน้ำหนักได้สูงสุด 150 กก.</d11>
            <d12>กำลัง: 2.0HP ความเร็ว 1-16 Km/h</d12>
            <d13>Console สามารถปรับได้ ด้วยไฮไลท์สีดำเพื่อการมองเห็นได้ชัดขึ้น</d13>
            <d14>
                การออกแบบปุ่มสำหรับการปรับความเร็วในการใช้งานที่ง่าย เพื่อตอบสนองความ
                ต้องการของผู้ใช้งานได้อย่างรวดเร็ว และแม่นยำ
            </d14>
            <d15>Stereo inside การออกกำลังกายที่มาพร้อมกับความบรรเทิง</d15>
            <d16>
                Shock-absorbing design ช่วยดูดซับแรงกระแทกจากการวิ่ง ทำให้เครื่องทำงาน
                เงียบ
            </d16>
            <d17>เพิ่มประสิทธิภาพการทำงานของระบบหัวใจและไหลเวียนของเลือด</d17>
```

```
            <d18>เพิ่มประสิทธิภาพการทำงานของระบบหายใจ</d18>
            <d19>
                ถ้าหากออกกำลังกายอย่างเป็นประจำ ก็จะสามารถลดน้ำหนักได้เป็นอย่างดี โดยเป็น
                การเพิ่มกระบวนการในการเผาผลาญไขมัน
            </d19>
            <d20>สามารถปรับระดับความชันของการวิ่งได้ถึง 15 ระดับ</d20>
            <d21>พับเก็บได้สะดวก ประหยัดพื้นที่</d21>
            <d22>เครื่องบริหาร SIX PACK CARE (ซิก แพค แคร์)</d22>
            <d23>
                เครื่องบริหารลดหน้าท้อง ถูกออกแบบมาเพื่อการบริหารหน้าท้องแบบ 2
                ทิศทาง และซิทอัพได้กว่า 180 องศา ช่วยให้การบริหารทำได้อย่างมีประสิทธิภาพ
                กว่าการบริหารบนพื้นถึง 2 เท่า มาพร้อมคุณสมบัติแบบ 6 in1 ในเครื่องเดียว ทั้ง
                บริหารได้ขณะเอนตัวลง และบริหารได้ขณะยกตัวขึ้น ที่เน้นการบริหารกล้ามเนื้อ
                หน้าท้อง , บริหารแบบทวิส ช่วยเน้นการบริหารกล้ามเนื้อข้างลำตัว, บริหาร
                กล้ามเนื้อหน้าท้องส่วนล่างและกล้ามเนื้อต้นขา ,บริหารร่างกายส่วนบนด้วยท่าวิด
                พื้น, บริหารร่างกายส่วนล่าง ขา และน่อง ครบทุกฟังก์ชั่นที่ช่วยให้คุณบริหารได้
                หลายสัดส่วนยิ่งขึ้น
            </d23>
            <d24>
                ท่าบริหารแบบซิทอัพ ช่วยให้การบริหารกล้ามเนื้อหน้าท้องทำได้ง่าย
                และมีประสิทธิภาพ
            </d24>
            <d25>ท่าบริหารแบบซิทอัพ (อย่างมืออาชีพ) เน้นการบริหารกล้ามเนื้อหน้าท้อง</d25>
            <d26>บริหารแบบทวิส ช่วยเน้นการบริหารกล้ามเนื้อข้างลำตัว</d26>
            <d27>สายแรงต้าน ใช้บริหารกล้ามเนื้อช่วงต้นแขน ไหล่ กล้ามเนื้อแขน</d27>
            <d28>
                บริหารร่างกายด้วยท่าวิดพื้น ออกกำลังกายท่อนบนอย่างเข้มข้นตั้งแต่ ไหล่ หน้าอก
                แขน ต้นแขน
            </d28>
            <d29>บริหารร่างกายส่วนล่าง ขา ต้นขา สะโพก รวมถึงกล้ามเนื้อน่อง</d29>
            <d30>รองรับน้ำหนักผู้ใช้งานสูงสุด 120 กก.</d30>
            <d31>ตัวเครื่องสามารถพับได้ทำให้จัดเก็บได้สะดวกยิ่งขึ้น</d31>
        </detail>
    </item>
    .
    .
    .
    <time_second>266.20422600</time_second>
</store>
```

รูปที่ ก.7 ตัวอย่างผลการทดลองจากเว็บไซต์ 360TVShopping

```
<store name="reebok" crawl_date="2014-12-14 14:38:36">
    <item>
        <subject>NPC II</subject>
        <url>http://global.reebok.com/5258_570.html</url>
        <detail>
            <d0>Back</d0>
            <d1>Home</d1>
            <d2>&gt; Men</d2>
            <d3>&gt; Shoes</d3>
            <d4>&gt;NPC II</d4>
            <d5>Men Classics</d5>
            <d6>ColorWhite/Lt Grey(5258)</d6>
            <d7>Store Locator</d7>
            <d8>Send to a Friend</d8>
            <d9>
                Anything but boring. A modern twist on a fashion icon , the NPC. Soft
                leather uppers for supportive comfort. High abrasion outsole provides
                durability. The ability to customize and extra laces in bold colors help
                you make leave a lasting impression.
            </d9>
            <d10>
                Leather upper / textile lining / synthetic sole
            </d10>
            <d11>Soft leather uppers/supportive comfort</d11>
        </detail>
    </item>
    .
    .
    .
    <time_second>1,439.19031700</time_second>
</store>
```

รูปที่ ก.8 ตัวอย่างผลการทดลองจากเว็บไซต์ Global Reebok

ภาคผนวก ข
ผลงานที่ได้รับการตีพิมพ์

Knowledge and Smart Technology (KST), 2015 7th International Conference, IEEE Transection.

Web Content Extraction Based on Subject Detection and Node Density

Warid Petprasit and Saichon Jaiyen

Department of Computer Science, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Thailand

s5650804@kmitl.ac.th, kjsaicho@kmitl.ac.th

Currently, very large data have been transferred from everywhere through World Wide Web. Consequently, the information extraction systems have been arising and many researches have been focusing on those data for utilizing them. These systems are very useful for data pre-processing and cleaning for real-time applications. Moreover, these systems can make other analyzing systems to analyze the data in real time such as social network mining, web mining, data mining, or even special tasks such as false advertisement detection, demand forecasting, and comment extraction on product and service reviews. In this paper, we focus on extracting the content data of web pages in e-commerce web sites based on subject detection and node density. In the experimental results, it can signify that our proposed method is appropriated to extract the data rich region in data-intensive pages in an automatic fashion.

Keywords- web information extraction; web content extraction; e-commerce; subject detection; node density (SDND); web mining; data intensive; wrapper induction.

I. INTRODUCTION

The information extraction is very useful for pre-processing the data in many fields such as web mining, recommendation system, decision making, expert system, knowledge discovery and so on. It is also useful to special tasks such as false advertisement detection, demand forecasting, and comment extraction on product reviews. There are a lot of researches that focus on these challenge problems. Since many web page structures rely on web developer and programming language, there are a lot of styles of placing content data on web pages. Consequently, it is very difficult to extract the content data automatically. Many researches focus on extracting the details of the content data from link offer page [3, 4, 5, 7, 10, 11, 12, 13]. Repetitive Subject Pattern (RSP) method manually receives the subject from the user as its input [3] after that it generate wrapper rule for extracting remain data in that pages. In addition, they extend RSP by improving the algorithm to add the subject automatically [4]. HyLiEn [5] and VIPS [7] use the vision based technique and the block model for webpage segmentation but it does not identify that which part is the content. DSE [13] proposes the method to improve performance of information extraction systems by removing noises before extracting. Some researches focus on annotating the output using regular expression [12] after extracting. Some researches focus on the data rich region at data intensive page [1, 6] because it has many descriptions of item. DE-SSE [1] uses structural-semantic entropy to detect the data rich region node. However, DE-SSE needs the user to define the keywords of the target node and then it calculates structural-semantic entropy value from annotated keywords. Since, it transforms the input keyword to regular expression firstly, this make DE-SSE depended on regular expression. Because most end users don't know about regular expression, this makes DE-SSE not suitable for end user. CECTD-DS [6] uses text density and link density for information extraction on the data-intensive pages from weblog. However, CECTD-DS cannot be applied to extract e-commerce web sites because some content nodes of e-commerce web sites have low text density. In this paper, we focus on extracting the content data on data-intensive pages of e-commerce web sites using the subject detection and the node density.

This paper is organized as follows. Section 2 mentions about the proposed method, how to identify the subject node location and define the data rich region node including pseudo code for implementing it. Section 3 describes why we choose those datasets, shows the experimental result of this research and the comparison between the proposed method and CECTD-DS. The last section of this paper mentions conclusion and future work.

II. SUBJECT DETECTION AND NODE DENSITY

A. Subject Detection

The web information extraction is becoming important for pre-processing the data before analyzing the data. There are several techniques to extract the content of web pages. Most approaches use semi-automatic techniques to extract the content data. These techniques need some help from human to identify some parameters. In this paper, we automatically detect the subject before extracting the content data on data-intensive pages of e-commerce web sites. The subject is the topic of data-intensive pages of e-commerce web sites. Normally, the subject of e-commerce web sites is a product name and the subject of weblogs is a topic name. In this paper, we propose the algorithm to identify the subject node that uses the tag name, the keywords in meta tag and title tag, and some properties in cascading style sheet (CSS) including font-weight, font-size, and display properties. The subject node can be found by finding the node that has the maximum total weight. In many web sites use <h1> and <h2> as a subject but for some web sites use other tags as a subject. Therefore, in this research, we assigned <h1> tag to have the highest weight and the weights of other tags will be decreased in descending

978-1-4799-6049-1/15/$31.00 ©2015 IEEE

order. Let S = {<h1>, <h2>, <h3>, , } be the set of candidate tags that may be the subject tag. These tags are usually used by search engine optimization (SEO) technique for ranking webpages. For each node in DOM tree whose tag name is the tag in S, the total weight of the *i*th node can be calculated by:

$$W_i = (\alpha \times N_i) + (\beta \times T_i) + (\gamma \times K_i) + (\delta \times C_i) \quad (1)$$

where W_i is the total weight of the *i*th candidate tag and N_i is the weight of the tag name of the *i*th candidate tag. The parameters including α, β, γ, and δ are a constant whose value is between 0 and 1 such that $\alpha + \beta + \gamma + \delta = 1$. These constants are used to normalize the values of total weight. In this paper, the weight of the tag name is calculated by:

$$N_i = \begin{cases} 1.00 & if\ name=<h1> \\ 0.75 & if\ name=<h2> \\ 0.50 & if\ name=<h3> \\ 0.25 & if\ name=<div>,<span> \end{cases} \quad (2)$$

Let T_i be the similarity between words in the title tag and words in the candidate tags. Let A be the set of all words contained in the title tag and B_i be the set of all words contained in the *i*th candidate tag. T_i can be calculated as:

$$T_i = \frac{|A \cap B_i|}{max(|A|, |B_i|)} \quad (3)$$

Let K_i be the similarity between words in the meta tag whose value of name attribute is "keywords" and words in the *i*th candidate tag. Let C be the set of all words in the meta tag whose value of the name attribute is "keywords" and D_i be the set of all words in the *i*th candidate tag. K_i can be computed as:

$$K_i = \frac{|C \cap D_i|}{max(|C|, |D_i|)}$$

Let C_i be the weight of CSS properties which is calculated from some CSS properties including display, font weight, and font size. The weight of CSS properties can be calculated as:

$$C_i = (\kappa \times display_i) + (\lambda \times fontweight_i) + (\mu \times fontsize_i) \quad (4)$$

where κ, λ, and μ are a constant whose value is between 0 and 1 such that $\kappa + \lambda + \mu = 1$. These constants are used to normalize the values of CSS properties. Let $display_i$ be the weight of the display property of CSS in the *i*th candidate tag. The weight of the display property can be computed by:

$$display_i = \begin{cases} 1 & if\ display = block \\ 0 & if\ otherwise \end{cases} \quad (5)$$

Let $fontweight_i$ be the weight of the font weight property in the *i*th candidate tag. The weight of the font weight property is calculated by:

$$fontweight_i = \begin{cases} 1 & if\ fontweight = bold, bolder \\ 0 & if\ otherwise \end{cases} \quad (6)$$

Let $fontsize_i$ is the weight of the font size property in the *i*th candidate tag. The weight of the font size property can be computed by:

$$fontsize_i = \frac{(fs_i - minfs)}{(maxfs - minfs)} \quad (7)$$

where fs_i is the value of the font- size property of the *i*th candidate tag , *minfs* is the minimum value of all font sizes in that document, *maxfs* is the maximum value of all font sizes in that document.

In this paper, we propose the algorithm to detect the subject node of data-intensive pages of e-commerce web sites. Firstly, the html source code is retrieved from the data intensive page. Secondly, the html source codes are parsed to the DOM Tree. Thirdly, for each tag in candidate tags, the total weight is calculated using equation (1). Finally, the node that has the highest total weight is assigned to be the subject node and send it to Algorithm 2. The proposed algorithm for automatically detecting the subject node is shown in Algorithm 1. In this paper, the adjustable weights are set as $\alpha = 0.5$, $\beta = 0.1$, $\gamma = 0.1$, $\delta = 0.3$, $\kappa = 0.15$, $\lambda = 0.15$, and $\mu = 0.70$. The tag name and CSS properties are important to identify the subject node. Thus, their weights are set higher than the others.

Algorithm 1: Subject Detection

Input: URL of data intensive page

Output: Subject node

1. Retrieve HTML Source Code
2. Parse HTML to DOM Tree
3. **for each** node *i* in DOM Tree **do**
4. **if** the tag name is an element of candidate tags **then**
5. Calculate total weight of node *i*
6. Find the node that has the maximum value of total weights and assign it to be a subject node.

B. Node Density

Node Density method is used to find the data rich region node before the extracting process. In e-commerce web site, the data rich region node is the node in DOM tree that contains the product detail or content data that keep only the needed information in that page. In this paper, we propose the new algorithm to identify the data rich region node which

contains the content data by defining link nodes as noise nodes which are mentioned in CECTD-DS [6]. The content data is the main detail of something such as the detail of the product. After the data rich region node is detected, the content data is extracted from this node. The input of this proposed algorithm is the output of Algorithm 1, which is the subject node. After taking its input, it assigns the current node as the subject node. Then, it reaches the parent node of the current node for deciding whether the node is the data rich region by using the threshold. This threshold can be calculated as:

$$threshold = \frac{node\ density - link\ density}{link\ density} \quad (8)$$

where the *node density* is number of all nodes in the current node and the *link density* is the number of link nodes (tag <a>).

Algorithm 2: Node Density

Input: Subject node

Output: XML file

1. Set the current node equal to the subject node and set the previous threshold as 0
2. Compute the current threshold using equation 8
3. **if** current threshold ≥ previous threshold **then**
4. Assign previous threshold equal to the current threshold
5. Go to the parent node of the current node and assign it to be the current node
6. Go to step 3
7. **else**
8. Assign the current node to be the data rich region node
9. Extract the detail to XML file

When the algorithm has found the data rich region node, it stops running and continues to extracting the detail or content of the current node. In this paper, the content of the data rich region node is extracted in the form of Extensible Markup Language (XML) file because it is easy to read or write by both humans and machines. However, the output of the proposed algorithm can be other types because it is extracted in the form of a sub tree. The proposed algorithm for identifying the data rich region node and extracting the content of this node is shown in Algorithm 2. Firstly, it receives the subject node from Algorithm 1 and assigns it to the current node. Then, the value of the starting threshold is set as 0. Secondly, the current threshold is calculated by using equation (8). Thirdly, the algorithm checks whether the current threshold is greater than or equal to the previous threshold. If the current threshold value is greater than or equal to the previous threshold, it means that the current node is not the data rich region node. Then, the current node is assigned as its parent node and go to step 2. Otherwise, the current node is the data rich region node. When the data rich region node is found, the content data is extracted from this node.

III. Experiment and Result

A. Datasets

In this paper, all dataset are collected from multiple sources for verifying the performance of the proposed method. These dataset are shown in Table I. There are 7 attributes in this table including datasets URL, extracted pages, language, size, objective, and actual shop. The dataset URL is the source of the dataset from which the pages are extracted. The extracted pages attribute shows the number of pages that are extracted for the experiment. The size attribute describes the number of data-intensive pages and the number of users that access to that web site. The objective attribute describes the characteristic of the websites which are built for sell only (B2C) and built for buy, sell or auction (C2C). The product type attribute is the category of the product in that websites. The actual shop attribute shows whether there exists the real shop that users can go to the shop which is shown in that web site. Most researches collect their own dataset because the data in web pages are updated and changed every rapidly in World Wide Web.

B. Experimental Results

In this paper, we apply precision, recall, and f-measure to verify the performance of the proposed algorithms. It can be calculated by using equation (9), (10), and (11) respectively.

Table I. Dataset.

Dataset URL	Extracted pages	Language	Size	Objective	Product type	Actual shop
http://www.hngcase.com/	150	Thai	Small	B2C	Mobile Case	No
http://www.nadzproject.com/?lang=th	150	Thai	Medium	B2C	Game	Yes
http://www.ksscom.co.th/	150	Thai	Medium	B2C	IT product	Yes
http://global.reebok.com/	150	English	Large	B2C	Reebok's brand	Yes
http://www.olx.co.th/?from=www	150	Thai	Large	C2C	All	No
http://www.asos.com/men/	150	English	Large	B2C	Fashion	No
https://www.jib.co.th/web/	150	Thai	Medium	B2C	IT product	Yes
http://www.360tvshopping.com/	147	Thai	Small	B2C	Fitness equipment	No

Table II. Experimental Results of proposed method.

Dataset	Extracted Page	Relevant Node	Correct Node	Incorrect Node	Recall	Precision	F-Measure
Hngcase	150	8250	8250	900	100.00%	90.16%	94.83%
Nadzproject	150	2550	2550	464	100.00%	84.61%	91.66%
Ksscom	150	3900	3768	0	96.62%	100.00%	98.28%
OLX	150	1500	1500	0	100.00%	100.00%	100.00%
ASOS	150	4050	3882	0	95.85%	100.00%	97.88%
JIB Computer	150	7050	6595	949	93.55%	87.42%	90.38%
360TVShopping	147	3969	3969	0	100.00%	100.00%	100.00%
Global Reebok	150	1350	1339	882	99.19%	60.29%	74.99%
Overall	**1197**	**32619**	**31853**	**3195**	**98.15%**	**90.31%**	**93.50%**

The precision define the accuracy of the proposed method. The recall shows the ability of data extraction of the proposed method. The f-measure is calculated based on the precision and recall.

$$Recall = \frac{correct\ node}{relevant\ node} \quad (9)$$

$$Precision = \frac{correct\ node}{correct\ node + incorrect\ node} \quad (10)$$

$$F\ measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (11)$$

In this experiment, we use several public tools for checking the results of this paper. Google Chrome browser is used for viewing XML file. Extensible Stylesheet Language (XSL) is used for querying the XML file (output of this method stored in XML). XSL programming language is used for querying the details of item from XML output file. The experimental results are shown in Table II. There are 8 datasets in this experiment. Each dataset consist of a lot of pages to be extracted. In addition, there are 8 attributes in this table. The datasets URL attribute shows the source from which the dataset are collected. The extracted page attribute shows the number of pages in that dataset to be extracted. The relevant node attribute shows the number of all relevant nodes in that page. The correct node is the extracted nodes that match with relevant nodes and the incorrect nodes is the extracted nodes that not match with relevant nodes. Precision, recall, and f-measure are already mentioned above. From the overall result shown in Table II, it can be seen that the proposed method give the precision, recall, and f-measure as 90.31%, 98.15%, and 93.50%, respectively. These can show that our proposed method is appropriated to extract the data rich region in data-intensive pages automatically.

Table III. The comparison between SDND and CECTD-DS.

Dataset	SDND F-Measure	CECTD-DS F-Measure
Hngcase	94.83%	93.04%
Nadzproject	91.66%	81.63%
Ksscom	98.28%	72.11%
OLX	100.00%	42.28%
ASOS	97.88%	81.41%
JIB Computer	90.38%	0.00%
360TVShopping	100.00%	62.07%
Global Reebok	74.99%	75.00%
Overall	**93.50%**	**63.44%**

Table III shows the comparison results between the proposed method (SDND) and CECTD-DS [6]. From these results, it can be seen that the proposed method can achieve the best result at 93.50% of f-measure. However, CECTD-DS is used for identifying the content block of webpages using text and link density. Unfortunately, the content block of e-commerce websites have the low text density. Consequently, CECTD-DS gives the low values of f-measure as shown in Table III. Thus, CECTD-DS is not suitable for e-commerce websites because its content block has the low text density.

IV. Conclusion

In this paper, we propose a new method to find the data rich region in data-intensive pages using subject detection and node density. It is every useful for pre-processing the data before analyzing by many techniques such as data mining, product analysis, decision support system, and recommendation system. It works by finding the subject location in the web page and then identifying the data rich region by traversing back to the parent of the subject node until the rate of link node is more than none link node. After that, we extract the received DOM Tree to XML file. The advantages of the proposed method are that it can easily be implemented because it does not require any input except only the URL of data-intensive pages and it can be simply used by end users. Some methods still require the expert user to identify the keyword and to define regular expression for it. Therefore, those methods are not suitable for end users. From the experimental results, it can be signified that our proposed method are appropriated to extract the data rich region in data-intensive pages automatically. It not needs the user to identify anything in the algorithm. Furthermore, the proposed algorithm can extract the data rich region in data-intensive pages in unsupervised

manner. However, the limitation of this proposed method is that it depends on subject detection method and link density in the data rich region. In the future work, we will extend this method to overcome this limitation and apply it to weblogs or news web sites.

REFERENCES

[1] X. Zheng, Y. Gu, and Y. Li, "Data Extraction from Web Pages Based on Structural-Semantic Entropy," International World Wide Web Conference Committee (W3C2). pp. 93-102, ACM, 2012.

[2] Brauer et al. "Enabling Information Extraction by Inference of Regular Expressions from Sample Entities," ACM International Conference on Information and Knowledge Management (CIKM). pp. 1285-1294, ACM, 2011.

[3] W. Thamviset, and S. Wongthanavasu, "Structured Web Information Extraction Using Repetitive Subject Pattern," Electrical Engineering/Electronics Computer Telecommunications and Information Technology (ECTICON). pp. 1-4, IEEE, 2012.

[4] W. Thamviset, and S. Wongthanavasu, "Information Extraction for Deep Web Using Repetitive Subject Pattern," Machine Learning and Intelligent System (MLIS), pp. 1109-1139, Springer, 2013.

[5] F. Fumarola, T. Weninger, and R. Barber, "Extracting General Lists from Web Documents: A Hybrid Approach (HyLiEn)," International Conference on Industrial Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE), pp. 285-294, ACM, 2011.

[6] F. Sun, D. Song, and L. Liao, "DOM Based Content Extraction via Text Density," Special Interest Group on Information Retrieval, ACM, 2011.

[7] D. Cai, S. Yu, J. Wen, and W. Ma, "VIPS: a Vision-based Page Segmentation Algorithm," Microsoft Research, pp. 1-29, 2003.

[8] C. Chang, M. Kayed, M. Girgis, and K. Shaalan, "A Survay of Web Information Extraction System," IEEE Transactions on Knowledge and Data Engineering, pp. 1411-1428, IEEE, 2006.

[9] A. Bhardwaj, and V. Mangat, "A Novel Approach for Content Extraction from Web Pages," Engineering and Computational Sciences (RAECS), pp. 1-4, IEEE, 2014.

[10] C. Chang, C. Hsu, and S. Lui, "IEPAD: Information Extraction Based on Pattern Discovery," Proceedings of the 10th international conference on World Wide Web, pp. 681–688, ACM,, 2001.

[11] H. Sleiman, and R. Conchuelo, "A Survey on Region Extractors from Web Documents," IEEE Transactions on Knowledge and Data Engineering, pp. 1960-1981, IEEE, 2013.

[12] I. Jellouli, and M. Mohajir, "Towards automatic semantic annotation of data rich Web pages," Research Challenges in Information Science, pp. 139-142, IEEE, 2009.

[13] J. Wang, and F. Loshovsky, "Data-rich Section Extraction from HTML pages," Web Information Systems Engineering (WISE), pp. 1-10, IEEE, 2002.

[14] Leander et al, "A Brief Survey of Web Data Extraction Tools," Special Interest Group on Management of Data (SIGMOD), pp. 84-93, ACM, 2002.

[15] M. Schreng, "Webbots Spiders and Screen Scrapers A Guild to Developing Internet Agents with PHP/CURL," No Starch Press, 2012.

International Joint Conference on Computer Science and Software Engineering (JCSSE), 2015 12th, IEEE Transection.

E-Commerce Web Page Classification Based on Automatic Content Extraction

Warid Petprasit and Saichon Jaiyen
Department of Computer Science, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Thailand
s5650804@kmitl.ac.th, kjsaicho@kmitl.ac.th

***Abstract*—Currently, There are many E-commerce websites around the internet world. These E-commerce websites can be categorized into many types which one of them is C2C (Customer to Customer) websites such as eBay and Amazon. The main objective of C2C websites is an online market place that everyone can buy or sell anything at any time. Since, there are a lot of products in the E-commerce websites and each product are classified into its category by human. It is very hard to define their categories in automatic manner when the data is very large. In this paper, we propose the method for classifying E-commerce web pages based on their product types. Firstly, we apply the proposed automatic content extraction to extract the contents of E-commerce web pages. Then, we apply the automatic key word extraction to select words from these extracted contents for generating the feature vectors that represent the E-commerce web pages. Finally, we apply the machine learning technique for classifying the E-commerce web pages based on their product types. The experimental results signify that our proposed method can classify the E-commerce web pages in automatic fashion.**

Keywords-SDND; machine learning; subject detection; node density; e-commerce; C2C; web mining; web page classification; product classification; product categorization; feature selection; markov random field; MRFs.

I. INTRODUCTION

Currently, web page classification tasks have usually been interested in many fields of research. K. Dong, L. Guo and Q. Fu developed detection method for detecting the pornographic image by using both of image context feature and image text feature [5] and then they apply Support Vector Machine (SVM) to classify the data. V.V. Sawant and V.R. Ghorpade proposed framework for classifying and categorizing web services by using generated semantic meta data [7]. W. Di, C. Wah, A. Bhardwaj, R. Piramuthu and N. Sundaresan presented fine-grain clothing style recognition and retrieval in fashion categorization by using automatic image tagging [19]. R.K. Rao, S.A. Khan, Z. Begum and C. Divakar classified the customer's behaviors to build E-commerce personalized system [9]. B.K. Mohanty and K. Passi applied fuzzy approach to classify the product properties and details from multiple search engines [21]. From these researches, we can see that various research fields in web mining have been interested in the web page classification. For many online stores, the product classification based on data mining techniques has been often used for product and service recommendation, fraud detection and false categorization. Since, the types of products have categorized by owner before register the product in C2C (Customer to Customer) web site. It is very hard to categorize products by people when there are tons of products available. For these reasons, an automatic product classification can help us to perform the automatic product categorization to reduce the burden of human.

Normal websites are easy to extract the content because they contain a lot of texts in the content, some models only use the text density to extract the content node. For E-commerce web pages, they contain the complicate structures such as layout, image, and multiple type of content (description, price, shipping, and warranty). It is very hard to identify which part is the content. In this research, we propose the method for classify E-commerce web pages based on their product types using Multi Layers Perceptron, automatic content extraction, and automatic keyword extraction. This paper is organized as follows. Section I mentions about the related works for web pages classification. Section II describes the related method used in this research. Section III describes the proposed method. Section IV describes the experiment results and parameter setting for each model used in the experiment. The last section summarizes entire paper and lead to some new ideas in the future.

II. RELATED WORKS

A. Subject Detection and Node Density

Subject Detection and Node Density (SDND) [22] are used in web content extraction task. Firstly, it finds the web subject location based on the observation that CSS properties of the subject have the biggest font size and the bold text. After detecting the subject node, it use the detected subject location as the starting point and crawling back to the parent node of the current node until the node density of the current node become unusual. Then, it assigns the previous node as the data rich region node or the content node. This method can be applied to prepare the data before performing the automatic web page classification. For this reason, we adopt SDND for extracting the content of E-commerce web pages in automatic manner.

B. Markov Random Field

Markov Random Field (MRFs) [20] is the feature selection technique for selecting the set of optimal features in order to reducing the number of features in the dataset. In addition, it can be applied to a small dataset with a large number of features. In this paper, we propose the automatic key word extraction method by adopting the MRFs to automatically extract the keywords from the contents of E-commerce web pages. N. Claypo and S. Saichon [15] have shown that the

978-1-4799-1966-6/15/$31.00 ©2015 IEEE

MRFs is suitable for selecting the key words from the web documents in order to perform the opinion mining for Thai restaurant reviews.

C. *Multi Layers Perceptron*

Multi Layers Perceptron (MLP) is popular method that is widely used in classification problems. It consists of input layer, hidden layer, and output layer. The MLP can be applied to learn the patterns of data for pattern classification. It can achieve high accuracy in the classification of text documents. Z. Chen, C. Ni and Y.L. Murphy presented text document categorization with neural network in a large dataset with multiple classes [13]. D. Saha mentioned that neural network gave the high accuracy when it was applied to a very large dataset [24].

D. *Naïve Bayes*

Naïve Bayes (NB) is the statistical model that apply the probability theory to predict class label of unseen data. It has been widely used in classification problems because it is easy to use and it can works well in a small dataset. R.D. Goyal mentioned that Naïve Bayes will give the high accuracy in the small size of vocabularies [12]. In addition, it was combined with neural network to improve the accuracy in the small dataset for text classification. B. Soiraya, A. Mingkhwan and C. Haruechaiyasak studied the factors that influence the design of E-Commerce web sites [6]. They show that Naïve Bayes can achieve the highest accuracy compared to Support Vector Machine based on Sequential Minimal Optimization (SMO).

E. *Radial Basis Function*

Radial Basis Function (RBF) neural network is another type of artificial neural networks that adopts the radial basis function as its kernel function in hidden layers. It uses a prototype to represent the center of the basis function and uses the spread parameter to represent the radius of the basis function. C.Y. Change and S.Y. Fu applied RBF and MRBF (Module Radial Basis Function) for texture image classification [11]. C. Junjie and H. Rongbing applied RBF and Back Propagation (BP) neural network to classify web pages from 1,000 pages. They showed that the RBF gave the higher accuracy than BF [17].

F. *Support Vector Machine*

Support Vector Machine (SVM) is one of classification models in machine learning. SVM classify the data based on the optimal hyperplane that is used to separate the data. The advantage of SVM for classifying text and web documents is that it can achieve the high accuracy when it is used to classify the high dimensional data [10, 14, 23]. R. Rajalakshmi and C. Aravindan applied SVM for classifying web pages. Y. Xi applied SVM to classify Chinese review spam [8]. H. Shi and X. Li used SVM to perform sentiment analysis in hotel reviews [3].

III. PROPOSED METHOD

In this paper, we propose the method for classifying E-commerce web pages based on automatic content extraction and automatic keyword extraction. Firstly, the SDND method is used to extract the contents of E-commerce web pages automatically. Secondly, the automatic keyword extraction based on MRFs is used to select the key words from the extracted contents to generate the input vectors that represent the web pages. Finally, the MLP is applied to classify the E-commerce web pages based on their categories.

A. *Automatic Content Extraction*

In this paper, the proposed SDND method [22] is applied to extract the contents of E-commerce web pages in automatic manner. The SDND consists of two stages. The first stage is the subject detection which is used to detect the subject node of E-commerce web pages. The four components of HTML including tag name, title tag, keyword tag, and CSS properties are used to calculate the weight of candidate tags for detecting the subject node. The candidate tag with the highest weight is selected as the subject node. The weight of each candidate tag can be calculated as:

$$W_i = (\alpha \times N_i) + (\beta \times T_i) + (\gamma \times K_i) + (\delta \times C_i) \quad (1)$$

where w_i is the weight of the ith candidate tag and α, β, γ, and δ is a constant such that $\alpha+\beta+\gamma+\delta=1$. N_i is the tag name of the i candidate tag whose value is *<h1>*, *<h2>*, *<h3>*, * *, or **. T_i is the similarity between the word in the title tag and the ith candidate tag. K_i is the similarity between words in the meta tag whose value of the name attribute is "keywords" and words in the ith candidate tag. C_i is the weight of CSS properties which is calculated from some CSS properties which are the display property, the font weight property, and the font size property. The weight of CSS properties can be computed as:

$$C_i = (\kappa \times display_i) + (\lambda \times fontweight_i) + (\mu \times fontsize_i) \quad (2)$$

where κ, λ, and μ are constant values such that $\kappa+\lambda+\mu = 1$.

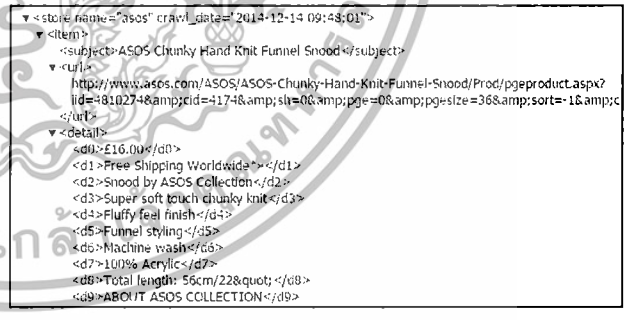


Figure I. The example of XML output from SDND

The second stage is the content node detection based on node density. After the subject node is detected, it is used as the starting node and it is used to crawl back to the parent node of the current node until it found the data rich region node. To identify the content node, it calculate the threshold according to equation (3) every time when it change the state

from the previous node to the current node. This threshold is used to identify the increasing rate of the link node between current node and previous node. When the increasing rate is unusual, it stop crawling and assign the previous node as the content node. This threshold can be calculated as:

$$threshold = \frac{node\,density - link\,density}{link\,density} \quad (3)$$

where the *node density* is number of all nodes in the current node and the *link density* is the number of link nodes (tag <*a*>). The output of SDND method is in the XML scheme as shown in figure I. It consists of *subject, url*, and *detail* tags.

B. Automatic Key Word Extraction

The text preparation technique is use to transform the text in documents into words and remove some irrelevant words from documents. It remove the stop words from all of sentences in documents such as *"a", "and", "the"*, etc. After the text preparation, the MRFs method is applied to select an optimal set of key words in order to generate the feature vectors and reduce the number of features in the dataset before applying machine learning techniques. This can be done in automatic manner. Finally, we classify the E-commerce web pages based on their categories using MLP neural network.

The proposed algorithm can be summarized as follows.

1. Input the E-commerce web page in HTML format.
2. Detect its subject and extract its content using SDND method.
3. Transform the text in the extracted content into words by using text preparation technique.
4. Apply MRFs method to select some related key words to generate the feature vectors.
5. Classify the E-commerce web pages using MLP neural network.

IV. Experimental Result

The dataset use in this paper are collected from 3 online store web sites including Jabong, Global Reebok, and Asos. These websites sell fashion and sport items in multiple categories. In this experiment, we choose 5 categories which are shoes, clothing, accessories, bag & purses, and beauty from these websites to generate the dataset. Each product detail in these websites is shown in a web page. We select 500 products and group them into 5 categories. Each class consists of 100 products and each product is detailed in a web page. The reason that we not use public dataset because the web sites have rapid growth and the web structure can be changed every time. Thus, when you find your product today, it will be disappeared or its structure will be changed in tomorrow because it sold out. The experiments are conducted 4 times using the dataset that is not preprocessed, the dataset that use only MRFs feature selection for preprocessing, the dataset that use only SDND method for preprocessing, and the dataset that use SDND and MRFs method for processing. All of them use the text preparation technique to cut all stop words.

In this paper, we adopt MLP neural network to classify these dataset and its performance is compared to Naïve Bay, RBF, and SVM. The 5-fold cross validation is used to evaluate all classification models. The 400 pages is used to train and 100 page is used to test. This process is repeated 5 times and the results are averaged. For SDND method, we set the constant values as $\alpha = 0.5$, $\beta = 0.1$, $\gamma = 0.1$, $\delta = 0.3$, $\kappa = 0.15$, $\lambda = 0.15$, and $\mu = 0.70$. These values are set as in the previous work [22]. For MRFs feature selection, the constant values are set as $\gamma = -0.5$ and $\beta = 0.03$. For MLP, the Scale Conjugate Gradient (SCG) [2] algorithm is used to train MLP with 16 hidden nodes. For Naïve Bayes, we use multinomial distribution that is frequently used in document classification problems. For RBF, the spread parameter is set as 0.5. For SVM, we apply SVM multiclass with Error-Correcting Output Codes (ECOC) [18] to classify the datasets. The experiments are conducted on the personal computer with CPU Intel Core i7-2670QM 2.20 GHz, and RAM 8.00 GB. The Operating System is Window 7 and the software used in the experiments is MATLAB.

The experiment results are shown in Table I. In the head of Table I, the number in a pair of parentheses is the number of selecting features after preprocessing. In this paper, we propose the method for E-commerce web page classification using MLP neural network with automatic content extraction and automatic keyword extraction. From Table I, it can be seen that our propose method can achieve the highest accuracy of 97.60 percent whereas Naïve Bay can achieve the accuracy of 78.20 percent, RBF can achieve the accuracy of 91.00 percent, and SVM can achieve the accuracy of 95.00 percent. For the dataset that is not preprocess, SVM can achieve the highest accuracy of 91.60 percent. For the dataset that use only SDND method for preprocessing, SVM can give the highest accuracy of 95.00 percent. For the dataset that use only MRFs method for preprocessing, SVM can give the highest accuracy of 89.60 percent. However, for the dataset that use SDND and MRFs method for preprocessing, MLP can

Table I. The comparison results.

Classifier	All features (8,307)		SDND (5,415)		MRFs (2,760)		SDND+MRFs (960)	
	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time
MLP	80.60	6.5343	83.20	4.9395	81.60	2.3004	**97.60**	1.3774
Naïve Bayes	63.40	0.6789	74.80	0.4502	58.20	0.2301	78.20	0.1861
RBF	80.40	**0.1591**	87.00	**0.0809**	83.40	**0.0670**	91.00	**0.0555**
SVM	**91.60**	1.1160	**95.00**	0.6212	**89.60**	0.3209	95.20	0.1894

achieve the highest accuracy of 97.60 percent. We can see that MLP and SVM give the high accuracy of classification because this dataset may be suitable for linear classifier models such as MLP and SVM. In addition, it can be seen that our proposed SDND technique can improve the performance of all classifier models and give the higher accuracy than the normal classification of dataset with no preprocessing. Furthermore, the proposed preprocessing method based on SDND and MRFs method can improve the accuracy of all classification model and it can reduce the training time of classification models. From these experimental results, it can signify that the proposed method is suitable for E-commerce web page classification based on product categories.

V. CONCLUSION

In this paper, we propose the method for classifying the E-commerce web pages using MLP neural network, automatic content extraction, and automatic key word extraction. We adopt our proposed SDND method to automatically extract the contents of E-commerce web pages. Then, we apply Markov Random Field (MRFs) to select the optimal set of keywords from the extracted contents automatically. Eventually, we apply the MLP neural network to classify the E-commerce web pages into their categories. From the experimental results, it can be seen that the proposed preprocessing method can improve the accuracy of all classification models. The performance of the proposed method is evaluated and compared to Naïve Bay, RBF, and SVM. From the experimental results, it can be seen that the proposed method can achieve the highest accuracy of 97.60 percent. These experimental results show that our propose method is appropriated for E-commerce web page classification. In the further work, we will extend SDND to extract the web content of other types of E-commerce web sites.

VI. REFERENCES

[1] X. Sun, Y.Liu, Y. Chai and H. Sun, "A Method for Online Retail Sales Estimation based on Semantic Features of Web Pages", 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp.236-241, IEEE, May, 2014.

[2] M. F. Moller, "A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning", Neural Networks, Vol 6, Issue 4, pp.525-533, ScienceDirect, Nov, 1993.

[3] H. Shi and X. Li, "A Sentiment Analysis Model for Hotel Reviews based on Supervised Learning", International Conference on Machine Learning and Cybernetics (ICMLC), pp.950-954, IEEE, Jul, 2011.

[4] G. Daqi and Y. Genxing, "Adaptive RBF Neural Networks for Pattern Classifications", International Joint Conference on Neural Networks (IJCNN), pp.846-851, IEEE, May, 2002.

[5] K. Dong, L. Guo and Q. Fu, "An Adult Image Detection Algorithm based on Bag-of-Visual-Words and Text Information", 10th International Conference on Natural Computation (ICNC), pp.557-560, IEEE, Aug, 2014.

[6] B. Soiraya, A. Mingkhwan and C. Haruechaiyasak, "An Analysis of Visual and Presentation Factors Influencing the Design of E-commerce Web Sites", International Conference on Web Intelligence and Intelligent Agent Technology (WIC), pp.525-528, IEEE, Dec, 2008.

[7] V. V. Sawant and V. R. Ghorpade, "Automatic Semantic Classification and Categorization of Web Services in Digital Environment", International Conference on Computer and Communications Technologies (ICCCT), pp.1-6, IEEE, Dec, 2014.

[8] Y. Xi, "Chinese Review Spam Classification Using Machine Learning Method", International Conference on Control Engineering and Communication Technology (ICCECT), pp.669-672, IEEE, Dec, 2012.

[9] R. K. Rao, S. A. Khan, Z. Begum and C. Divakar, "Design of E-Commerce Personalized Service Model based on Web Mining Classification Technique", International Conference on Computational Intelligence and Computing Research (ICCIC), pp.1-4, IEEE, Dec, 2013.

[10] M. S. Othman, L. M. Yasuf and J. Salim, "Features Discovery for Web Classification Using Support Vector Machine" , International Conference on intelligent Computing and Cognitive Informatics (ICICCI), pp.36-40, IEEE, Jun, 2010 .

[11] C. Y. Chang and S. Y. Fu, "Image Classification using a Module RBF Neural Network", International Conference on Innovative Computing, Information and Control (ICICIC), pp.270-273, IEEE, Sep, 2006.

[12] R. D. Goyal, "Knowledge Based Neural Network for Text Classification", International Conference on Granular Computing (GRC), pp.542-547, IEEE, Nov, 2007.

[13] Z. Chen, C. Ni and Y. L. Murphey, "Neural Network Approaches for Text Document Categorization", Internation Joint Conference on Neural Networks (IJCNN), pp.1054-1060, IEEE, 2006.

[14] P. Sahoo, A. K. Behera and M. K. Pandia, "On the Study of GRBF and Polynomial Kernel based Support Vector Machine in Web Logs", International Conference on Emerging Trends and Applications in Computer Science (ICETACS), pp.1-5, IEEE, Sep, 2013.

[15] N. Claypo and S. Jaiyen, "Opinion Mining for Thai Restaurant Reviews using K-Means Clustering and MRF Feature Selection", International Conference on Knowledge and Smart Technology (KST), pp.105-108, IEEE, Jan, 2015.

[16] R. Patel and P. Thakkar, "Opinion Spam Detection Using Feature Selection", International Conference on Computational Intelligence and Communication Networks (CICN), pp.560-564, IEEE, Nov, 2014.

[17] C. junjie and H. Rongbing, "Research of Web Classification Mining based on RBF Neural Network", Control, Automaton, Robotics and Vision Conference (ICARCA), pp.1365-1367, IEEE, Dec, 2004.

[18] T. G. Dietterich and G. Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes", Journal of Artificial Intelligence Research, pp.263-286, ACM, Aug, 1994.

[19] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu and N. Sundaresan, "Style Finder: Fine-Grained Clothing Style Detection and Retrieval", Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.8-13, IEEE, Jun, 2013.

[20] Q. Cheng, H. Zhou and J. Cheng, "The Fisher-Markov Selector: Fast Selecting Maximally Separable Feature Subset for Multiclass Classification with Applications to High-Dimentionnal Data", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.1217-1233, ACM, Jun, 2011.

[21] B. K. Mohanty and K. Passi, "Web based Information for Product Ranking in E-Business - A Fuzzy Approach", International Conference on Electronic Commerce (ICEC), pp.558-563, ACM, Aug, 2006.

[22] W. Petprasit and S. Jaiyen, "Web Content Extraction based on Subject Detection and Node Density", International Conference on Knowledge and Smart Technology (KST), pp.121-125, IEEE, Jan, 2015.

[23] R. Rajalakshmi and C. Aravindan, "Web Page Classification using n-gram based URL Features", International Conference on Advanced Computing (ICoAC), pp.15-21, IEEE, Dec, 2013.

[24] D. Saha, "Web Text Classification Using a Neural Network", International Conference on Emerging Applications of Information Technology (EAIT), pp.57-60, IEEE, Feb, 2011.

[25] J. Han, M. Kamber and J. Pei, "Data Mining Concepts and Techniques", Third Edition, Elsevier, 2012.

ประวัติผู้เขียน

ชื่อ	นายวริษฐ์ เพ็ชรประสิทธิ์
วัน เดือน ปีเกิด	8 มีนาคม พ.ศ. 2533
ที่อยู่	บ้านเลขที่ 407/33 ถ.มิตรภาพ ต.ปากเพรียว อ.เมือง จ.สระบุรี
ประวัติการศึกษา	2554 บริหารธุรกิจบัณฑิต สาขาระบบสารสนเทศทางคอมพิวเตอร์-คอมพิวเตอร์ธุรกิจ เกรดเฉลี่ย 3.11 มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี ปทุมธานี
ผลงานทางวิชาการ	เรื่อง การสกัดรายละเอียดผลิตภัณฑ์จากหน้าเว็บโดยวิธีการตรวจหาโหนดหัวข้อและความหนาแน่นของโหนด (WEB CONTENT EXTRACTION BASED ON SUBJECT DETECTION AND NODE DENSITY) 7th International Conference on Knowledge and Smart Technology (KST) ระหว่างวันที่ 28-31 เดือนมกราคม 2558 ณ เดอะไทด์ รีสอร์ท (หาดบางแสน) จังหวัดชลบุรี