ขั้นตอนและวิธีการสกัดข้อมูลสินค้าบนเว็บเพจ สำหรับเว็บครอเลอร์ที่ใช้ในโปรแกรมค้นหาสินค้า

An Algorithm of Product Information Extraction on Web Pages for Web Crawler in Product Search Engines

กลยุทธ บพิตร

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตาม
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมเว็บ
คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2555

ขั้นตอนและวิธีการสกัดข้อมูลสินค้าบนเว็บเพจ สำหรับเว็บครอเลอร์ที่ใช้ในโปรแกรมค้นหาสินค้า

กลยุทธ บพิตร

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตาม หลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมเว็บ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2555

An algorithm of product information extraction on web pages for web crawler in product search engines

Kollayuth Borpit

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Web Engineering

Faculty of Information Technology, Dhurakij Pundit University

2012

กิตติกรรมประกาศ

ผู้จัดทำสารนิพนธ์ขอกราบคารวะ และขอบูชาพระคุณครูบาอาจารย์ คณะเทคโนโลยีสารสนเทศ สาขาวิศวกรรมเว็บ มหาวิทยาลัยธุรกิจบัณฑิตย์ทุกท่านที่ได้เสียสละ เพื่อให้ศิษย์ทุกคนมีวิชาความรู้ติดตัว และสามารถนำความรู้ที่ได้มาใช้ในการปฏิบัติงานจริงในอนาคตได้

สารนิพนธ์ฉบับนี้ สำเร็จลงได้ด้วยความอนุเคราะห์จาก ผศ.ดร.วรสิทธิ์ ชูชัยวัฒนา ผู้จัดทำสารนิพนธ์ใคร่ขอกราบขอบพระคุณที่ท่านอาจารย์กรุณาให้คำปรึกษา ให้ข้อคิดเห็น และเสียสละเวลาให้กับผู้จัดทำสารนิพนธ์ตลอดมา ตลอดจนช่วยตรวจสอบต้นฉบับและแก้ไขข้อบกพร่องของงานวิจัยเพื่อให้สารนิพนธ์ฉบับนี้เสร็จสิ้นลงไปได้ด้วยดี ซึ่งส่งผลให้สารนิพนธ์ฉบันนี้ดำเนินการได้อย่างราบรื่น และมีประสิทธิภาพ

ขอกราบขอบพระคุณท่านอีกครั้งเพราะนอกจากการช่วยเหลือด้านการเรียนแล้วท่านยังให้การช่วยเหลือเมื่อมีปัญหาในทุกเรื่อง พร้อมทั้งแนะแนวทางการดำเนินชีวิตด้วยการให้คำปรึกษาที่ดีมาโดยตลอด

ขอกราบขอบพระคุณ รศ.ดร.นุชรี เปรมชัยสวัสดิ์, และ ดร.ธนภัทร์ อนุศาสน์อมรกุลที่กรุณารับเป็นกรรมการสอบสารนิพนธ์ และกรุณาให้คำแนะนำที่มีค่าอย่างยิ่ง

สุดท้ายนี้ ผู้จัดทำสารนิพนธ์ ขอกราบขอบพระคุณ คุณพ่อ คุณแม่ ที่ส่งเสียเล่าเรียน และขอบคุณเพื่อนๆ ทุกคนที่เป็นกำลังใจอันสำคัญยิ่งในการทำสารนิพนธ์ ซึ่งจะถูกจารึกไว้ในจิตใจของผู้ทำสารนิพนธ์ด้วยความระลึกถึงตลอดไป

กลยุทธ บพิตร

สารบัญ

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

ภาพที่ **หน้า**

หัวข้อสารนิพนธ์	ขั้นตอนและวิธีการสกัดข้อมูลสินค้าบนเว็บเพจสำหรับเว็บครอเลอร์ที่ใช้ในโปรแกรมค้นหาสินค้า
ชื่อผู้เขียน	นายกลยุทธ บพิตร
อาจารย์ที่ปรึกษา	ผศ.ดร.วรสิทธิ์ ชูชัยวัฒนา
สาขาวิชา	วิศวกรรมเว็บ
ปีการศึกษา	2554

บทคัดย่อ

สารนิพนธ์นี้ได้นำเสนอขั้นตอนและวิธีการสกัดข้อมูลสินค้าสำหรับเว็บคอรเลอร์ที่ใช้ในโปรแกรมค้นหาสินค้า โดยเน้นการวิเคราะห์โครงสร้างของเอกสาร และการวิเคราะห์คำสำคัญ ขั้นตอนและวิธีการที่เสนอ เริ่มต้นจากการวิเคราะห์โครงสร้างหน้ารวมของเว็บขายสินค้า จากนั้นทำการค้นหาโหนดของสินค้าโดยการวิเคราะห์รูปและราคา และนำโหนดของสินค้าที่ได้ไปทำการสกัดรายละเอียดของข้อมูลสินค้า

สำหรับการประเมินประสิทธิผลของการทำงานของขั้นตอนและวิธีการสกัดข้อมูลสินค้า ผลลัพธ์ที่ได้จากการทำงานของขั้นตอนวิธีการสกัดข้อมูลสินค้าที่เสนอ จะถูกเปรียบเทียบกับผลลัพธ์การสกัดข้อมูลสินค้าโดยมนุษย์ ในการทดลองผู้วิจัยทำการสุ่มเลือกยูอาร์แอลจากชนิดของเว็บไซต์พาณิชย์อิเล็กทรอนิกส์ 3 ประเภทได้แก่ เว็บร้านค้าออนไลน์ เว็บไซต์แคตตาล็อกสินค้าออนไลน์ และเว็บตลาดกลางอิเล็กทรอนิกส์ จำนวนทั้งสิ้น 60 ยูอาร์แอล ผลการประเมินพบว่าขั้นตอนและวิธีการสกัดข้อมูลสินค้าและวิธีการสกัดข้อมูลสินค้าที่เสนอมีความแม่นยำในการสกัดข้อมูลสินค้าร้อยละ 88.4 สำหรับเว็บร้านค้าออนไลน์และเว็บแคตตาล็อกสินค้าออนไลน์ ในขณะที่มีความแม่นยำในการสกัดข้อมูลสินค้าร้อยละ 77.3 สำหรับเว็บตลาดกลางอิเล็กทรอนิกส์

อย่างไรก็ดีขั้นตอนและวิธีการสกัดข้อมูลสินค้าที่นำเสนอยังคงมีปัญหาในการทำงานในบางสถานการณ์เนื่องจากวิธีการดังกล่าวให้ความสำคัญกับการระบุตำแหน่งรูปภาพสินค้าก่อน ส่งผลทำให้วิธีการที่เสนอไม่สามารถทำการสกัดข้อมูลสินค้าที่ไม่มีรูปภาพได้ นอกจากนั้นแล้วการวิเคราะห์ราคาผิดพลาดก็เป็นอีกสาเหตุที่สำคัญที่ส่งผลต่อประสิทธิผลของการทำงานของขั้นตอนและวิธีการสกัดข้อมูลสินค้าที่นำเสนอ

Thematic Paper Title	An algorithm of product information extraction on web pages for web crawler in product search engines
Author	Mr. Kollayuth Borpit
Thesis Advisor	Asst. Prof. Dr.Worasit Choochaiwattana
Department	Web Engineering
Academic Year	2011

ABSTRACT

This thematic aims at proposing an algorithm of product information extraction on web pages for web crawler in product search engine focused on web document structure analysis and keyword analysis. The proposed algorithm starts with analyzing web document structure. The algorithm, then, finds product information nodes by analyzing product image and price. Finally, the product information is extracted.

For evaluating the proposed algorithm, the results obtained from the proposed algorithm is compare with the results from manually extraction by human. In the experiment, URLs were randomly selected from three categories URLs list, which are E-Shop website, E-catalogue website, and E-Marketplace website. The total number of URLs is 60 URLs. The results showed that the proposed algorithm had an accuracy in extracting product information of 88.4% for E-shop website and E-catalogue website, while the proposed algorithm had an accuracy of 77.3% for E-Marketplace website.

The proposed algorithm, however, still had problems in some situations because the proposed algorithm focused on locating product images first. As a result, the proposed algorithm could not extract product information on web pages that did not have product images. In addition, the problem in finding product price information is another main cause that affect an effectiveness of the proposed algorithm.

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของงาน

เนื่องจากในปัจจุบันการซื้อขายสินค้าผ่านระบบอินเทอร์เน็ตในประเทศไทยนั้นมีการขยายตัวมากขึ้น และพฤติกรรมการซื้อสินค้าออนไลน์กลายเป็นส่วนหนึ่งในชีวิตประจำวันของคนรุ่นใหม่ ในขณะเดียวกันมีการเพิ่มขึ้นของเว็บไซต์ในประเทศไทยโดยเฉพาะเว็บพาณิชย์อิเล็กทรอนิกส์ ส่งผลให้การค้นหาสินค้าที่ตรงกับความต้องการนั้นกลายเป็นภารกิจหนึ่งของผู้ที่สนใจซื้อสินค้าออนไลน์ ในปัจจุบันผู้ใช้งานอินเทอร์เน็ตในประเทศไทย มีการใช้โปรแกรมค้นหาข้อมูล (search engine) เช่น www.google.com สูงถึง 99.2% (ศูนย์วิจัยนวัตกรรมอินเทอร์เน็ตไทย, 2555) แต่ผลลัพธ์การค้นหาที่ได้ ยังคงไม่ตรงกับความต้องการของผู้ใช้งานที่ต้องการค้นหาข้อมูลสินค้า ดังนั้นทำให้ผู้ใช้งานยังคงต้องเข้าไปดูผลลัพธ์การค้นหา เพื่อพิจารณาว่าเนื้อหาในเว็บไซต์ที่เป็นผลลัพธ์นั้น ตรงตามความต้องการของตนเองหรือไม่ เนื่องจากมีหลายเว็บไซต์ที่นำมาแสดงในผลลัพธ์การค้นหา เป็นเว็บที่ให้ข้อมูลสินค้า หรือเว็บข่าวเกี่ยวกับสินค้า หรือคำวิจารณ์สินค้า ไม่ใช่เว็บสำหรับขายสินค้าโดยตรง ส่งผลให้ผู้ใช้งานต้องเสียเวลาในการค้นหาเว็บขายสินค้า ดังนั้นจึงมีการพัฒนาโปรแกรมค้นหาสินค้า (Product Search Engine) เพื่ออำนวยความสะดวกให้กับผู้ใช้งานและทำให้ได้ผลลัพธ์ที่ตรงกับความต้องการมากขึ้น

อย่างไรก็ดี จากการศึกษาเว็บไซต์ priceza.com ซึ่งเป็นเว็บไซต์ที่ทำงานใกล้เคียงกับโปรแกรมค้นหาสินค้ามากที่สุดที่มีอยู่ในประเทศไทย พบว่าให้ผลการค้นหาที่ไม่ตรงกับความต้องการ เนื่องจากมีการนำเอาข้อมูลสินค้าจากร้านค้าที่ไม่ได้เปิดดำเนินการบนอินเทอร์เน็ตเข้ามารวมอยู่ในผลลัพธ์การค้นหาด้วย จากการทดลองใส่คำค้นหา “Apple iPhone 4S 64GB” ลงใน priceza.com พบว่า ให้ผลลัพธ์เพียง 2 รายการ เมื่อเปรียบเทียบราคาสินค้าจากในเว็บพบว่ามาจากเว็บขายสินค้าเพียง 3 เว็บไซต์และ 1 ร้านค้าเท่านั้น ทำให้ผลลัพธ์การค้นหาของเว็บนี้ไม่ครอบคลุมเว็บขายสินค้าที่มีมากมายในประเทศไทย

ดังนั้นเพื่อทำให้โปรแกรมค้นหาสินค้ามีความครอบคลุมของข้อมูลมากยิ่งขึ้น การเพิ่มความสามารถให้กับเว็บครอเลอร์ด้วยการวิเคราะห์แยกส่วนข้อมูลที่จำเป็นเช่น ชื่อสินค้า รูปภาพ

ราคา และรายละเอียด เพื่อลดภาระในการทำงานของส่วนการจัดทำดัชนีข้อมูล จะส่งผลให้โปรแกรมค้นหาสินค้าทำงานได้อย่างรวดเร็วและเป็นอัตโนมัติมากขึ้น

ดังนั้นงานวิจัยชิ้นนี้จึงเสนอขั้นตอนและวิธีการสกัดข้อมูลสินค้าสำหรับเว็บครอเลอร์เพื่อทำการเก็บรวบรวมข้อมูลสินค้า และนำไปใช้ในโปรแกรมค้นหาสินค้า

1.2 วัตถุประสงค์ของการศึกษา

1.2.1 เพื่อออกแบบขั้นตอนและวิธีการสกัดข้อมูลสินค้าสำหรับเว็บครอเลอร์ เพื่อวิเคราะห์แยกส่วนข้อมูลที่จำเป็นเกี่ยวกับสินค้าเช่น รูปภาพ ราคา รายละเอียดย่อ และยูอาร์แอลสินค้า สำหรับเว็บขายสินค้าในประเทศไทย

1.2.2 เพื่อวัดประสิทธิผลของขั้นตอนและวิธีการสกัดข้อมูลสินค้า ตามที่ได้ออกแบบในงานวิจัยนี้

1.3 ประโยชน์และผลที่คาดว่าจะได้รับ

1.3.1 เพื่อเป็นแนวทางในการสกัดข้อมูลสินค้าให้กับเว็บครอเลอร์สำหรับโปรแกรมค้นหาสินค้า

1.3.2 เพื่อเป็นแนวทางในการสกัดข้อมูลสำหรับโปรแกรมรูปแบบอื่นๆ

1.4 ขอบเขตการศึกษา/ข้อตกลงเบื้องต้นของการศึกษา

1.4.1 หน้าเว็บขายสินค้าที่นำมาทำการทดลองนั้น คือหน้ารวมสินค้าที่ภายในหนึ่งหน้าเว็บมีสินค้ามากกว่า 1 สินค้าและแต่ละสินค้ามีรูปภาพแสดงเพียงหนึ่งรูป และรูปภาพสินค้าต้องมีขนาดกว้างและยาวกว่า 50 พิกเซล

1.4.2 สินค้าแต่ละสินค้าแยกกันอยู่ภายใน DOM ของภาษา HTML

1.4.3 ใช้สกัดข้อมูลโดยมนุษย์เป็นเกณฑ์ในการวัดประสิทธิผลของวิธีการ

1.4.4 เว็บขายสินค้าที่ใช้ในการวิจัยเป็นเว็บในรูปแบบของ เว็บร้านค้าออนไลน์ (E-Shop Web Site) และเว็บตลาดกลางอิเล็กทรอนิกส์ (E-Marketplace)

1.4.5 งานวิจัยนี้คำนึงถึงประสิทธิผลของข้อมูลสินค้าที่ได้ทำการสกัดออกมา โดยไม่นำปัจจัยที่เกี่ยวข้องกับการวัดประสิทธิภาพ เช่น ความเร็วของการสกัดข้อมูล มาพิจารณา

1.5 นิยามศัพท์

ในสารนิพนธ์ฉบับนี้ได้นิยามศัพท์ที่ใช้ในการวิจัยไว้ ดังนี้

เว็บครอเลอร์ (WEB CRAWLER) มีชื่อเรียกหลายชื่อเช่น Robot Spider เว็บครอเลอร์ คือ โปรแกรมหนึ่งในโปรแกรมค้นหาที่มีหน้าที่ในการไปเก็บข้อมูลจากยูอาร์แอลต่างๆ ซึ่งมีวัตถุประสงค์เพื่อเก็บรวบรวมข้อมูลเพิ่มหรืออัพเดทข้อมูลที่มีอยู่แล้วในโปรแกรมค้นหา สำหรับโปรแกรมค้นหาสินค้า เว็บครอเลอร์มีหน้าที่ไปเก็บข้อมูลสินค้าจากเว็บขายสินค้า

บทที่ 2
วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดและทฤษฎีที่เกี่ยวข้อง

2.1.1 โปรแกรมค้นหา (search engine)

เนื้อหาในหัวข้อนี้อ้างอิงจาก (อรวรรณ บึงพรรัตน์, 2549, กันยายน)

หลักการของโปรแกรมค้นหา เริ่มจากเว็บครอเลอร์รวบรวมเว็บเพจจากอินเตอร์เน็ตมาทำดัชนีให้อยู่ในรูปแบบที่ง่ายต่อการสืบค้นข้อมูล เมื่อผู้ใช้ทำการบอกสิ่งที่ต้องการกับโปรแกรมค้นหา โปรแกรมจะสืบค้นและแนะนำแหล่งข้อมูลที่คาดว่าตรงกับความต้องการมากที่สุดให้แก่ผู้ใช้

2.1.1.1 ส่วนประกอบของโปรแกรมค้นหา

1) เว็บครอเลอร์ (Web crawler) คือ โปรแกรมไปเก็บรวบรวมข้อมูลเว็บเพจตามลำดับคิวที่ได้มีการจัดไว้ ลงในฐานข้อมูลขนาดใหญ่ ซึ่งจะกล่าวถึงรายละเอียดในหัวข้อ 2.1.2

2) ส่วนดัชนี (Indexer) คือ โปรแกรมที่ทำหน้าที่สร้างดัชนีจากเอกสาร HTML ที่อยู่ในฐานข้อมูล ที่ Web crawler รวบรวมมา

การทำดัชนีมีวัตถุประสงค์ 3 ข้อในการค้นคืนสารสนเทศคือ

- ทำให้การหาที่อยู่ของเอกสารตามหัวเรื่องได้โดยง่ายและรวดเร็ว
- นิยามสาขาของหัวเรื่องและความสัมพันธ์ของเอกสารอันหนึ่งกับอีกอันหนึ่ง
- ทำนายความเกี่ยวข้องของเอกสารที่กำหนดเข้ากับความต้องการของข่าวสารที่ระบุ

3) หน้าเว็บค้นหา (Searcher) คือ โปรแกรมที่รับคำค้นหาของผู้ใช้งานซึ่งเป็นสิ่งที่ผู้ใช้ต้องการ เข้าสู่ระบบ จากนั้นก็จะทำการเปรียบเทียบคำค้นหากับดัชนีที่ได้สร้างไว้ และส่งเป็นผลลัพธ์การค้นหากลับไป

2.1.1.2 ชนิดของโปรแกรมค้นหา

1) Crawler Based Search Engines

Crawler Based Search Engines คือ โปรแกรมค้นหาบนอินเตอร์เน็ตแบบอาศัยการบันทึกข้อมูล และจัดเก็บข้อมูลเป็นหลัก ซึ่งเป็นโปรแกรมค้นหาที่ได้รับความนิยมสูงสุดเนื่องจากให้ผลการค้นหาแม่นยำที่สุด และการประมวลผลการค้นหาสามารถทำได้อย่างรวดเร็ว จึงทำให้มีบทบาทในการค้นหาข้อมูลมากที่สุดในปัจจุบัน

Crawler Based Search Engines มีองค์ประกอบหลัก 2 ส่วน ส่วนแรกคือ ฐานข้อมูล โดยส่วนใหญ่แล้ว Crawler Based Search Engine เหล่านี้จะมีฐานข้อมูลเป็นของตัวเองที่มีระบบการประมวลผล และการจัดอันดับที่เฉพาะเป็นเอกลักษณ์ของตนเองอย่างมาก และส่วนที่สองคือ ซอฟต์แวร์ คือ เครื่องมือหลักสำคัญที่สุดอีกส่วนหนึ่งของโปรแกรมค้นหาประเภทนี้คือ เว็บครอเอลร์ หรือที่เรียกว่า Spider หรือ Search Engine Robots ซึ่งสามารถแบ่งชนิดของเว็บครอเอลร์ได้เป็น 2 ชนิดคือ

(1) Horizontal crawler หรือ Breadth-first search ซึ่งทำงานด้วยการเก็บรวบรวมข้อมูลหน้าเว็บทั้งหมด โดยไม่กำหนดหัวเรื่องที่ต้องการเก็บ

(2) Vertical crawler หรือ focus crawler หรือ topical crawler มีการทำงานด้วยการเก็บรวบรวมข้อมูลเฉพาะที่ตรงประเด็นกับหัวข้อที่ได้กำหนดไว้ก่อนแล้วเท่านั้น

Crawler Based Search Engine ได้แก่ Google , Yahoo, MSN, Live, Search เป็นต้น

2) Web Directory หรือ Blog Directory

Web Directory หรือ Blog Directory คือ สารบัญเว็บไซต์ที่สามารถค้นหาข่าวสารข้อมูลด้วยหมวดหมู่ข่าวสารข้อมูลที่เกี่ยวข้องกัน ในปริมาณมากๆ คล้ายกับสมุดหน้าเหลือง ซึ่งจะมีการสร้าง ดรรชนี มีการระบุหมวดหมู่ อย่างชัดเจนซึ่งจะช่วยให้การค้นหาข้อมูลต่างๆ ตามหมวดหมู่นั้นๆ ได้รับการเปรียบเทียบอ้างอิงเพื่อหาข้อเท็จจริงได้ในขณะที่เราค้นหาข้อมูล เพราะว่าจะมีเว็บไซต์ หรือ Blog จำนวนมากที่มีเนื้อหาลักษณะเดียวกันในหมวดหมู่เดียวกัน ให้เลือกหาข้อมูลได้อย่างตรงประเด็นที่สุดตัวอย่างโปรแกรมค้นหาชนิดนี้ได้แก่

(1) ODP หรือ Dmoz คือ Web Directory ที่ใหญ่ที่สุดในโลก Search Engine หลายแห่งก็ใช้ข้อมูลจากที่แห่งนี้เกือบทั้งสิ้น เช่น Google, AOL, Yahoo, Netscape และอื่นๆ อีกมากมาย ODP มีการบันทึกข้อมูลประมาณ 80 ภาษาทั่วโลกรวมถึงภาษาไทย (URL : http://www.dmoz.org)

(2) สารบัญเว็บไทย SANOOK.com คือ Web Directory ที่มีชื่อเสียงอีกเช่นกัน และเป็นที่รู้จักมากที่สุดในเมืองไทย (URL: http://webindex.sanook.com)

3) Meta Search Engine

Meta Search Engine คือโปรแกรมค้นหาที่ใช้หลักการในการค้นหาโดยอาศัย Meta Tag ในภาษา HTML ซึ่งมีการประกาศชุดคำสั่งต่างๆ เป็นรูปแบบของ Tex Editor ด้วยภาษา HTMLเช่น ชื่อผู้พัฒนา คำค้นหา เจ้าของเว็บ หรือ บล็อกคำอธิบายเว็บหรือบล็อกอย่างย่อ

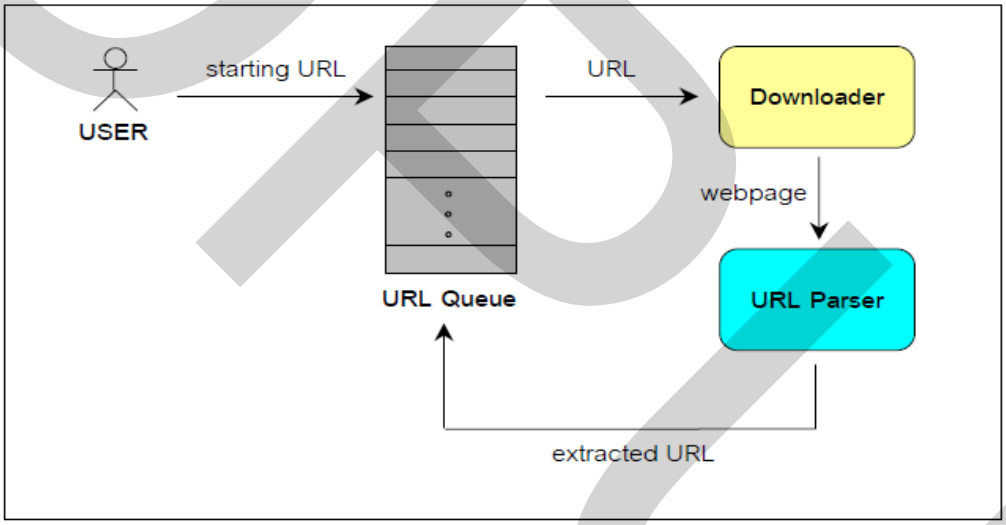
ผลการค้นหาของ Meta Search Engine นี้ มักไม่แม่นยำเนื่องจากบางครั้งผู้ให้บริการหรือผู้ออกแบบเว็บสามารถใส่อะไรเข้าไปได้มากมาย เพื่อให้เกิดการค้นหาและพบเว็บหรือบล็อกของตนเอง

2.1.2 ความรู้เบื้องต้นเกี่ยวกับเว็บครอเลอร์

เนื้อหาในหัวข้อนี้อ้างอิงจาก (นิรันดร์ อังควัฒนวิทย์, 2546: 3-5)

เว็บครอเลอร์เป็นโปรแกรมที่พัฒนาขึ้นเพื่อใช้ประโยชน์ในการรวบรวมเว็บเพจจากอินเทอร์เน็ต กลุ่มวิจัยหลากหลายกลุ่มต่างพยายามออกแบบและสร้างเว็บครอเลอร์ให้สามารถทำงานได้อย่างมีประสิทธิภาพสูงสุดอย่างไรก็ตามการออกแบบและลักษณะการทำงานยังคงยึดอยู่บนพื้นฐานเดียวกัน และปรับปรุงเปลี่ยนแปลงมาจากแบบจำลองพื้นฐานแบบเดียวกัน

2.1.2.1 แบบจำลองพื้นฐานของเว็บครอเลอร์



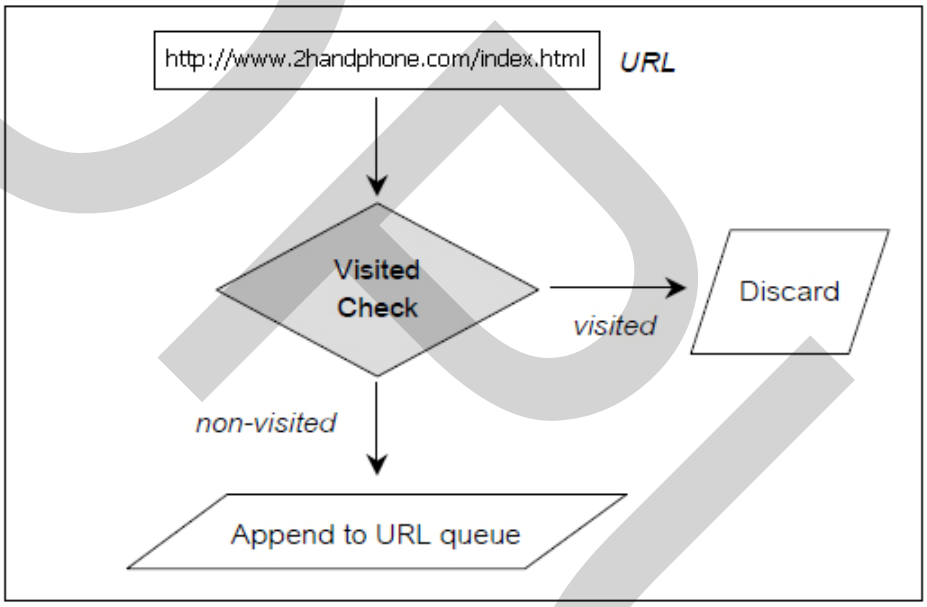
ภาพที่ 2.1 แบบจำลองพื้นฐานของเว็บครอเลอร์

ที่มา: นิรันดร์ อังควัฒนวิทย์ (2546: 3)

เว็บครอเลอร์เป็นโปรแกรมที่สามารถรวบรวมเว็บเพจได้โดยอัตโนมัติ โดยเริ่มเก็บเว็บเพจจากยูอาร์แอลเริ่มต้น จากนั้นยูอาร์แอลใหม่ที่อยู่ในเว็บเพจนั้นจะถูกแยกออกมาและนำไปตรวจสอบว่าก่อนหน้านี้ได้เก็บเว็บเพจจากยูอาร์แอลนี้มาก่อนหรือไม่ หากตรวจสอบพบว่าเคยเก็บมาก่อนจะไม่เก็บซ้ำซ้อนอีก แต่หากพบว่าไม่เคยเก็บมาก่อนเว็บครอเลอร์จะต้องเก็บเว็บเพจจากยูอาร์แอลนั้น ขั้นตอนเหล่านี้จะถูกทำงานวนรอบเสมอจนกว่าจะถึงจุดสิ้นสุดการทำงานซึ่งมี 2 กรณี คือ เก็บเว็บเพจได้ครบตามจำนวนที่กำหนด หรือไม่พบยูอาร์แอลที่สามารถเก็บต่อไปได้

แบบจำลองพื้นฐานของเว็บครอเลอร์ประกอบด้วย 3 ส่วนหลัก ได้แก่ ยูอาร์แอลคิว (URL queue) ตัวดาวน์โหลด (Downloader) และตัวพาร์สเซอร์ยูอาร์แอล (URL Parser) ดังภาพที่ 1 โดยแต่ละส่วนประกอบจะมีลักษณะดังนี้

1) ยูอาร์แอลคิว ทำหน้าที่เก็บยูอาร์แอลที่เว็บครอเลอร์พบในเว็บเพจ แต่ยังไม่ได้ไปเก็บรวบรวมมา มีลักษณะเป็นยูอาร์แอลคิวแบบเข้าก่อนออกก่อน ยูอาร์แอลคิวนี้จะมีกลไกในการตรวจสอบว่า ยูอาร์แอลที่เข้ามาเป็นยูอาร์แอลที่เว็บครอเลอร์เคยเก็บมาก่อนหรือไม่ หากเคยเก็บมาแล้วยูอาร์แอลนั้นจะไม่ถูกนำมาใส่ในยูอาร์แอลคิว ทั้งนี้เพื่อป้องกันความซ้ำซ้อนในการรวบรวมเว็บเพจ โดยกลไกของยูอาร์แอลคิวนี้แสดงดังภาพที่ 2.2



ภาพที่ 2.2 กลไกการตรวจสอบยูอาร์แอลก่อนนำไปใส่ในยูอาร์แอลคิว

ที่มา: นิรันดร์ อังควัฒนวิทย์ (2546: 4)

2) ตัวดาวน์โหลด ทำหน้าที่เก็บเว็บเพจจากอินเทอร์เน็ต โดยตัวดาวน์โหลดจะดึงยูอาร์แอลออกจากยูอาร์แอลคิว

3) ตัวพาร์สเซอร์ยูอาร์แอล ทำหน้าที่แยกส่วนของยูอาร์แอลที่พบในเว็บเพจออกมา โดยยูอาร์แอลที่ปรากฏในเว็บเพจจะถูกกำกับด้วยแท็กของภาษาเอชทีเอ็มแอล (HTML) ตัวพาร์สเซอร์ยูอาร์แอลจะอ่านเว็บเพจและแยกส่วนของยูอาร์แอลที่อยู่ภายในแท็ก <a href=URL > ออกมาเพื่อนำไปใส่ในยูอาร์แอลคิวต่อไป

2.1.2.2 การหายูอาร์แอลเริ่มต้น เทคนิคที่ใช้ในการหายูอาร์แอลเริ่มต้นมี 3 วิธี

1) ผู้ใช้กำหนดเอง (User Defined) วิธีนี้มักพบในเว็บครอเลอร์ที่ถูกสร้างขึ้นเป็นระบบใหญ่ โดยกำหนดให้ผู้ใช้ใส่ยูอาร์แอลเริ่มต้นด้วยตนเอง การให้ผู้ใช้สามารถกำหนดยูอาร์แอลเริ่มต้นเองได้เป็นสิ่งที่ดี เนื่องจากผู้ใช้อาจมีความเชี่ยวชาญในหัวเรื่องนั้นอยู่ก่อนแล้ว เว็บ ครอเลอร์จึงอาจได้ยูอาร์แอลเริ่มต้นที่ดีที่สุด อย่างไรก็ตามหากผู้ใช้คนหนึ่งไม่เชี่ยวชาญในหัวเรื่อง ทำให้การกำหนดยูอาร์แอลเริ่มต้นเองโดยผู้ใช้มีข้อด้อยมากกว่าข้อดี

2) การใช้เสิร์จเอ็นจิ้น (Search Engine) วิธีนี้เป็นการหายูอาร์แอลเริ่มต้นแบบอัตโนมัติโดยการนำคำสำคัญของหัวเรื่องที่กำหนดให้ไปสืบค้นกับเสิร์จเอ็นจิ้น จากนั้นนำผลลัพธ์จากการสืบค้นมาเป็นยูอาร์แอลเริ่มต้น โดยวิธีนี้ไม่ต้องใช้ผู้เชี่ยวชาญเข้าไปเกี่ยวข้องกับกระบวนการเก็บรวบรวมเว็บเพจ อย่างไรก็ตามประสิทธิภาพของเสิร์จเอ็นจิ้นต้องมีผลกระทบต่อยูอาร์แอลเริ่มต้นอย่างหลีกเลี่ยงไม่ได้ หากผลลัพธ์จากการสืบค้นผิดพลาด ยูอาร์แอลเริ่มต้นเหล่านั้นอาจนำเว็บครอเลอร์ไปในเส้นทางที่ไม่ดีก็เป็นได้

3) ไดเร็กทอรี่ ที่สร้างโดยมนุษย์ผู้เชี่ยวชาญ (Expert human-edited Directory) ซึ่งไดเร็กทอรี่ คือกลุ่มของเว็บเพจที่ถูกจัดหมวดหมู่ตามหัวเรื่องที่กำหนด ภายในหัวเรื่องหนึ่งจะประกอบไปด้วยเว็บเพจที่กล่าวถึงเรื่องนั้น วิธีนี้เป็นการใช้เว็บเพจที่ถูกจัดหมวดหมู่แล้ว นำมาเป็นยูอาร์แอลเริ่มต้นให้แก่เว็บครอเลอร์ ไดเร็กทอรี่สามารถสร้างได้ 2 แบบคือ ใช้ระบบอัตโนมัติ และใช้มนุษย์ผู้เชี่ยวชาญ แน่นอนว่าไดเร็กทอรี่ที่สร้างโดยระบบอัตโนมัติมักมีความถูกต้องน้อยกว่าไดเร็กทอรี่ที่สร้างด้วยผู้เชี่ยวชาญ ดังนั้นการหายูอาร์แอลเริ่มต้นจากไดเร็กทอรี่แบบสร้างเองโดยมนุษย์ย่อมได้ยูอาร์แอลเริ่มต้นที่มีความเกี่ยวข้องกับหัวเรื่องอย่างแน่นอน ตัวอย่างไดเร็กทอรี่ ได้แก่ DMOZ ในประเทศไทยได้แก่ Truehits Directory, Sanook Directory, Narak Directory เป็นต้น

2.1.3 โปรแกรมค้นหาสินค้า (Product Search Engine)

โปรแกรมค้นหาสินค้า คือ โปรแกรมค้นหาชนิดหนึ่งที่มีความจำเพาะเจาะจงไปที่ข้อมูลสินค้าเท่านั้น ซึ่งแตกต่างจากโปรแกรมค้นหาธรรมดาที่เก็บข้อมูลทุกรูปแบบจากทุกเว็บไซต์ และการแสดงผลก็ไม่ได้แสดงแค่เพียงชื่อเว็บกับยูอาร์แอลเท่านั้น แต่มีการแสดงรูปภาพสินค้า ยูอาร์-แอลสินค้า ราคาสินค้า และรายละเอียดโดยย่อของสินค้า ตัวอย่างโปรแกรมค้นหาสินค้าได้แก่ Google shopping, Bing shopping, Yahoo shopping และเว็บในประเทศไทยที่มีรูปแบบใกล้เคียงกับโปรแกรมค้นหาสินค้ามากที่สุดได้แก่ Priceza.com

2.1.4 เว็บไซต์ Google.com

เว็บไซต์ Google (www.Google.com) เป็นเว็บไซต์ที่ให้บริการในการค้นหาข้อมูลบนอินเตอร์เน็ต โดยค้นหาข้อมูลจากข้อความหรือตัวอักษรที่พิมพ์เข้าไป แล้วทำการค้นหาข้อมูล

รูปภาพ หรือเว็บเพจที่เกี่ยวข้องนำมาแสดงผล เว็บไซต์ Google ได้รับความนิยมอย่างมากในกลุ่มผู้ใช้งานอินเทอร์เน็ตที่ต้องการค้นหาข้อมูล เว็บไซต์ Google แบ่งหมวดหมู่ของการค้นหาออกเป็น 4 หมวดหมู่ด้วยกันดังตารางที่ 2.1

ตารางที่ 2.1 หมวดหมู่การค้นหาของเว็บไซต์ Google.com

เว็บ (Web)	เป็นการค้นหาข้อมูลในรูปแบบของเว็บไซต์ต่างๆ ทั่วโลก โดยการแสดงผลจะแสดงเว็บไซต์ที่มีคำที่เป็น Keyword อยู่ภายเว็บไซต์นั้น
รูปภาพ (Images)	เป็นการค้นหารูปภาพจากการแปลคำ Keyword
กลุ่มข่าว (News)	เป็นการค้นหาข้อมูลที่เป็นเนื้อหาที่อยู่ในข่าวซึ่งมีการระบุชื่อผู้เขียนข่าว หัวข้อข่าว วันที่และเวลาที่โพสต์ข่าว
สารบบเว็บ (Web Directory)	Google มีการจัดประเภทของเว็บไซต์ออกเป็นหมวดหมู่ซึ่งเราสามารถค้นหาเว็บในเรื่องที่ต้องการตามหมวดหมู่ที่มีอยู่แล้วได้เลย

2.1.5 เว็บไซต์ Priceza.com

คือ เว็บไซต์ที่ให้บริการค้นหาข้อมูลราคาสินค้า ที่เก็บรวบรวมข้อมูลจากร้านค้า ทั้งจากร้านค้าที่มีหน้าร้านจริง (ร้านแบบออฟไลน์) เช่น Central, Power buy, Index living mall, Home pro, Big c, Jay mart, It city, World camera เป็นต้นและร้านค้าออนไลน์หรือเว็บพาณิชย์อิเล็กทรอนิกส์รายใหญ่ เช่น tarad.com, shoppingpping.com, shopping.co.th, shopat7.com, trendyday.com เป็นต้น(Priceza.com, 2551)

2.1.6 รูปแบบของเว็บขายสินค้า E-Commerce

ภาวุธ พงษ์วิทยภานุ (2551) รูปแบบของการทำเว็บไซต์ E-Commerce มีหลายประเภทตามรูปแบบการทำงานของเว็บไซต์ดังนี้

2.1.6.1 การประกาศซื้อ-ขาย (E-Classified)

เป็นรูปแบบเว็บไซต์ E-Commerce ที่เปิดโอกาสให้ผู้ที่สนใจประกาศความต้องการซื้อ-ขาย สินค้าของตนได้ภายในเว็บไซต์ โดยเว็บไซต์จะทำหน้าที่เสมือนกระดานข่าวและตัวกลางในการแสดงข้อมูลสินค้าต่างๆ และหากมีคนสนใจสินค้าที่ประกาศไว้ ก็สามารถติดต่อตรงไปยังผู้ประกาศได้ทันทีจากข้อมูลที่ประกาศอยู่ภายในเว็บไซต์ โดยส่วนใหญ่จะมีการแบ่งหมวดหมู่ของประเภทสินค้าเอาไว้ เพื่อให้ง่ายต่อการเข้าไปเลือกซื้อ-ขายสินค้าในเว็บไซต์ เช่น http://www.thaisecondhand.com/ การซื้อขายรูปแบบนี้ ผู้ขายไม่จำเป็นต้องมีเว็บไซต์ของตัวเอง แต่

ทำได้โดยอาศัยพื้นที่ของเว็บที่เปิดโอกาสให้ประกาศขายของ ก็สามารถเริ่มต้นการค้าขายได้แล้ว ข้อดี คือ เริ่มต้นได้ง่ายทันที และฟรี ข้อเสีย คือ ไม่เหมาะกับผู้ที่มีสินค้าเป็นจำนวนมากๆ

2.1.6.2 เว็บไซต์แคตตาล็อคสินค้าออนไลน์ (Online Catalog Web Site)

เป็นรูปแบบจัดทำเว็บไซต์ E-Commerce ในรูปแบบแคตตาล็อคออนไลน์ ที่มีรูปภาพและรายละเอียดสินค้าพร้อมที่อยู่เบอร์โทรติดต่อ ไม่มีระบบการชำระเงินผ่านทางเว็บไซต์ หรือตะกร้าสินค้าออนไลน์ ถ้าลูกค้าสนใจสามารถโทรสอบถามและสั่งซื้อสินค้าได้ ซึ่งเป็นการใช้เว็บไซต์ เหมือนเป็นโบร์ชัวร์หรือแคตตาล็อคออนไลน์ เพื่อให้ลูกค้าสามารถเข้ามาเลือกดูรายละเอียดสินค้าและราคาได้ จากทั่วประเทศหรือทั่วโลกผ่านทางเว็บไซต์ ข้อดีของเว็บแบบนี้คือสร้างได้ง่ายเหมาะกับการค้าในพื้นที่หรือประเทศเดียวกัน ข้อเสียคือ ไม่สามารถขายและรับเงินได้ทันทีจากลูกค้าที่ต้องการชำระเงินผ่านเว็บไซต์

2.1.6.3 ร้านค้าออนไลน์ (E-Shop Web Site)

เป็นรูปแบบเว็บไซต์ E-Commerce สมบูรณ์แบบ ที่มีทั้งระบบการจัดการสินค้า ระบบตะกร้าสินค้า (Shopping Cart) ระบบการชำระเงิน รวมถึงการขนส่งสินค้าครบสมบูรณ์แบบ ทำให้ผู้ซื้อสามารถสั่งซื้อสินค้าและทำการชำระเงินผ่านเว็บไซต์ได้ทันที โดยการชำระเงินส่วนใหญ่เป็นการชำระผ่านบัตรเครดิต

ในการจัดทำเว็บไซต์ลักษณะนี้ จำเป็นต้องมีระบบหลายๆ อย่างประกอบอยู่ภายใน ซึ่งมีความซับซ้อนและมีรายละเอียดในการจัดทำค่อนข้างมาก อย่างไรก็ตาม ในขณะนี้มีเว็บไซต์ E-Commerce สำเร็จรูป ที่พร้อมให้บริการและมีการทำงานที่สมบูรณ์ ทำให้สามารถเริ่มต้นทำเว็บลักษณะนี้ได้อย่างรวดเร็ว หากท่านสนใจ ร้านค้าออนไลน์ สามารถสมัครใช้บริการฟรี

2.1.6.4 การประมูลสินค้า (Auction)

เป็นเว็บไซต์ E-Commerce ที่มีรูปแบบของการนำสินค้ามาประมูลขาย โดยการแข่งขันเสนอราคาสินค้า หากผู้ใดเสนอราคาสินค้าได้สูงสุดในช่วงเวลาที่กำหนด ก็จะชนะการประมูลและสามารถซื้อสินค้าชิ้นนั้นไปได้ ด้วยราคาที่ได้กำหนดไว้ โดยส่วนใหญ่สินค้าที่นำมาประมูล หากเป็นสินค้าใหม่ ซึ่งหลังการประมูลสินค้าจะมีราคาที่ไม่สูงกว่าราคาท้องตลาด ยกเว้นสินค้าเก่าบางประเภท หากยิ่งเก่ามากยิ่งมีราคาสูง เช่น ของเก่า ของสะสม เป็นต้น เช่น http://auction.tarad.com/ และ http://www.ebay.com/

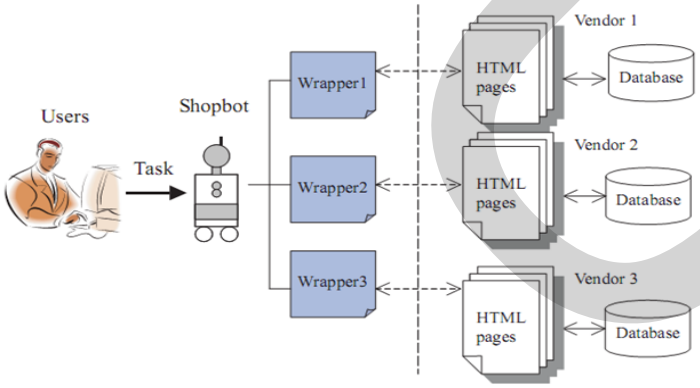
2.1.6.5 ตลาดกลางอิเล็กทรอนิกส์ (E-Marketplace)

เป็นเว็บไซต์ E-Commerce ที่มีรูปแบบเป็นตลาดนัดขนาดใหญ่ โดยภายในเว็บไซต์จะมีการรวบรวมเว็บไซต์ของร้านค้าและบริษัทต่างๆ มากมาย โดยมีการแบ่งหมวดหมู่ของสินค้าเอาไว้เพื่อให้ผู้ใช้สามารถเข้าไปดูสินค้าภายในร้านค้าต่างๆ ภายในตลาดได้ง่ายและสะดวก โดยรูปแบบ

ของตลาดกลางอิเล็กทรอนิกส์ บางแห่งมีการแบ่งออกเป็นหลายรูปแบบ ตามลักษณะของสินค้าที่มีอยู่ภายในตลาดแห่งนั้น เช่น ตลาดสินค้าทั่วไป http://www.tarad.com/ เว็บไซต์ตลาดกลางอิเล็กทรอนิกส์เกี่ยวกับอาหาร http://www.foodmarketexchange.com/ เว็บไซต์ตลาดกลางอิเล็กทรอนิกส์ของสินค้า OTOP อย่าง http://www.thaitambon.com/ เป็นต้น

2.2 งานวิจัยที่เกี่ยวข้อง

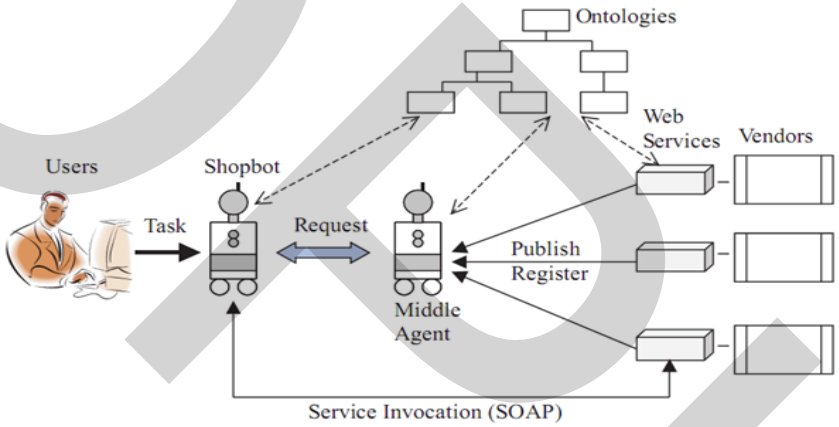
งานวิจัยในด้านนี้ปรากฏให้เห็นในกลางปี 1990 และนักวิเคราะห์เทคโนโลยีในขณะนั้นได้มีการทำนายว่า จะมีผลกระทบอย่างมากต่อผู้ขายและวิธีการดำเนินธุรกิจในอนาคต ตัวอย่างเว็บครอเลอร์สำหรับโปรแกรมค้นหาสินค้าในยุคแรกเรียกว่า Shopbot ซึ่งใช้สำหรับการเปรียบเทียบราคาคือ BargainFinder พัฒนาโดย (B.T. Krulwich, 1996: 258-263) ซึ่งอนุญาตให้ผู้ใช้สามารถเปรียบเทียบราคาของซีดีเพลงจากร้านค้าปลีกออนไลน์ แต่ร้านค้าปลีกจำนวนมากเริ่มบล็อกการเข้าถึงของเว็บครอเลอร์จาก BargainFinder เพราะเว็บครอเลอร์จาก bargainfinder นั้นมีการประเมินผู้ขายจากราคาจากราคาเท่านั้นและละเว้นคุณสมบัติอื่นๆ ทั้งหมดที่ร้านค้าปลีกเพลงออนไลน์ได้สร้างไว้ในเว็บไซต์ของตน ในที่สุด BargainFinder ต้องหยุดปฏิบัติการ



ภาพที่ 2.3 การทำงานของ Shopbot (R.B. Doorenbos, O. Etzioni, and D.S.Weld)

ที่มา: R.B. Doorenbos, O. Etzioni, and D.S.Weld.(1997: 43)

จากปัญหาที่ BargainFinder พบทำให้มีการพัฒนา (R.B. Doorenbos, O. Etzioni, and D.S.Weld., 1997:39-48) ซึ่งมีการทำงานดังภาพที่ 2.3 โดยเว็บครอเลอร์ที่ต้องทำงานผ่านตัวกลางที่เรียกว่า wrapper ซึ่งมีหน้าที่ในการสกัดข้อมูล ราคา รายละเอียดสินค้า หรือข้อมูลที่มักใช้ในการตัดสินใจซื้อสินค้า ออกมาจากเว็บไซต์แต่ละเว็บด้วยวิธี “screen-scraping” คือเข้าไปอ่านข้อมูลจากหน้าเว็บไซต์ที่ต้องการแล้วสกัดเอาข้อมูลและแท็กภาษา HTML ที่ไม่ต้องการทิ้งไป ให้เหลือเฉพาะข้อมูลที่ต้องการวิธีนี้ค่อนข้างยุ่งยาก ถ้าเว็บไซต์ต้นทางมีการเปลี่ยนโครงสร้างก็จะต้องแก้ไขโปรแกรมตาม และนอกจากนี้ wrapper แต่ละตัวทำงานได้กับเว็บเพียง 1 เว็บ Shopbot นี้ถูกเปลี่ยนชื่อในเวลาต่อมาเป็น jango และถูกซื้อและจดทะเบียนธุรกิจโดย excite ในปี 1997



ภาพที่ 2.4 การทำงานของ Shopbot ใน (Maria Fasli)

ที่มา: Maria Fasli.(2006: 72)

ต่อมาได้มีการพัฒนาเว็บครอเลอร์ที่ไปติดต่อกับส่วนจัดการข้อมูลของเว็บขายสินค้าโดยตรงคือ (Maria Fasli, 2006:69-75) มีการทำงานดังรูปภาพที่ 2.4 โดยมีการนำเทคโนโลยีใหม่ๆ เข้ามาช่วยได้แก่ เว็บเซอร์วิส (web service) และใช้ทฤษฎีในด้านออนโทโลยีในการจัดแบ่งหมวดหมู่สินค้า เว็บครอเลอร์จะทำงานโดยการร้องขอข้อมูลไปที่เว็บเซอร์วิสของเว็บขายสินค้าโดยตรงทำให้ได้ข้อมูลที่ถูกต้องแม่นยำและประหยัดเวลา แต่เนื่องจากงานวิจัยชิ้นนี้ทำขึ้นเพื่อโปรแกรมค้นหาสินค้าในประเทศไทยซึ่งมีเพียงไม่กี่เว็บเท่านั้นที่ใช้เทคโนโลยีเว็บเซอร์วิส วิธีนี้จึงไม่เหมาะที่จะนำมาใช้งาน

ดังนั้นงานวิจัยนี้จึงออกแบบและพัฒนาเว็บครอเลอร์ให้มีความสามารถในการคัดแยกข้อมูลสินค้าบางอย่างเช่น ราคา รูปภาพ ชื่อสินค้า ลิงค์สินค้า และรายละเอียดโดยไม่ผ่านตัวกลางหรือ wrapper (R.B. Doorenbos, O. Etzioni, and D.S.Weld., 1997: 39-48) และเก็บข้อมูลจากเว็บขายสินค้าในประเทศไทยให้ได้อย่างมีประสิทธิภาพ

บทที่ 3
วิธีการดำเนินการและเครื่องมือ

การวิจัยนี้เป็นการวิจัยเพื่อหาขั้นตอนและวิธีการสกัดข้อมูลสินค้าบนเว็บเพจสำหรับเว็บครอเลอร์ เพื่อนำไปใช้ในโปรแกรมค้นหาสินค้า

3.1 การศึกษาค้นคว้าข้อมูล

3.2 การออกแบบขั้นตอนและวิธีการสกัดข้อมูลสินค้า

3.3 การกำหนดแบบแผนการวัดประสิทธิผล

3.4 การวัดประสิทธิผล

3.5 เครื่องมือที่ใช้ในการวิจัย

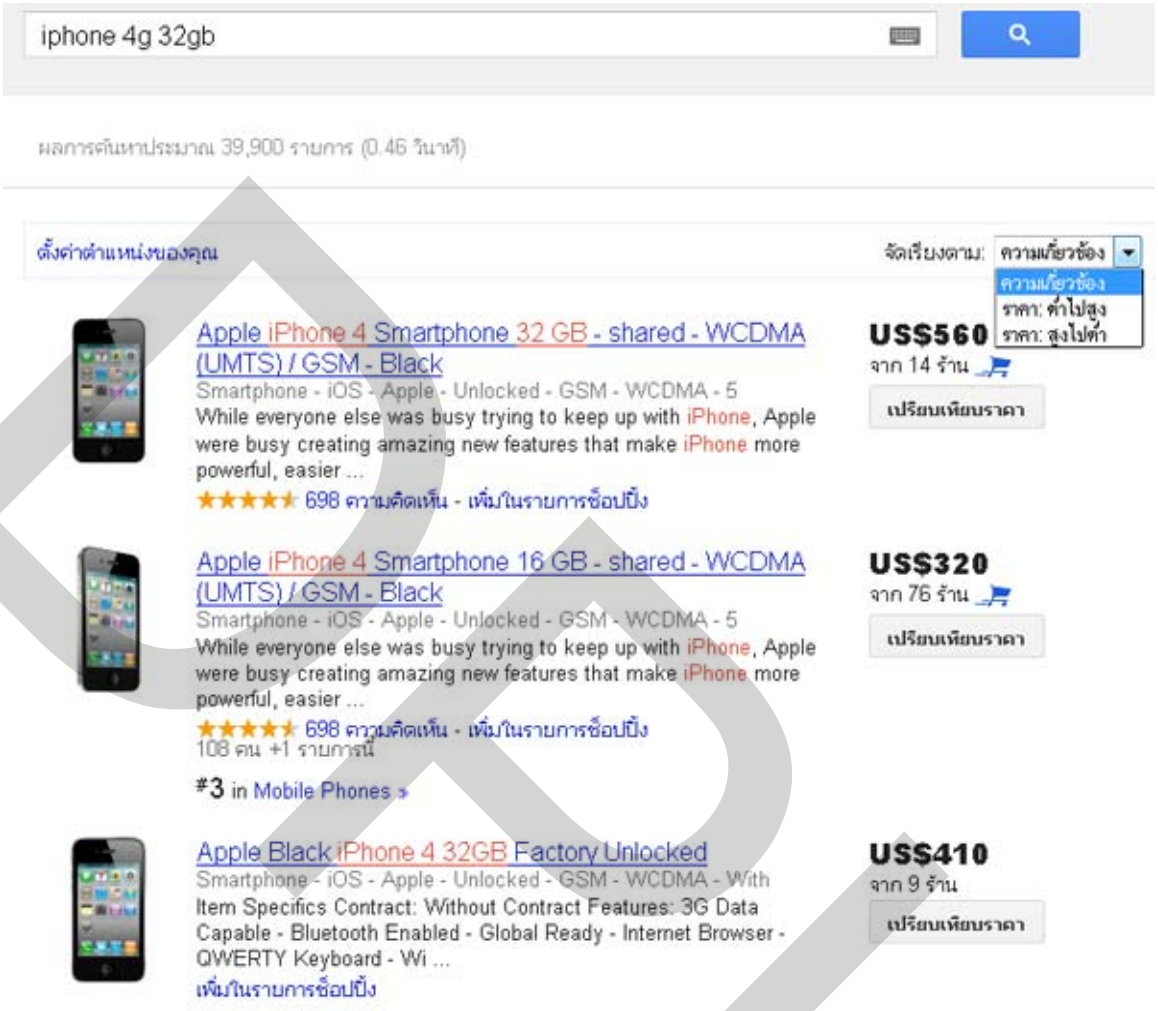
3.1 การศึกษาค้นคว้าข้อมูล

ผู้วิจัยได้ศึกษาและค้นคว้าข้อมูลจากแหล่งต่างๆ ในหัวข้อต่อไปนี้

3.1.1 ศึกษาเกี่ยวกับโปรแกรมค้นหาสินค้า

งานวิจัยนี้ได้ทำการศึกษาโปรแกรมค้นหาที่มีอยู่ในปัจจุบันและกำลังเป็นที่นิยม ทั้งในต่างประเทศเช่น Google Product Search, Shopping.com, Price Grabber และ Shopzilla และของไทยเช่น priceza.com ซึ่งพบว่า ข้อมูลที่สำคัญและจำเป็นที่สุดสำหรับโปรแกรมค้นหาสินค้าได้แก่ ชื่อสินค้า รูปภาพสินค้า ราคาสินค้า ยูอาร์แอลสินค้า และรายละเอียดย่อของสินค้าดังภาพที่ 3.1

จากการศึกษาการใช้งานโปรแกรมค้นหาสินค้า พบว่าข้อมูลข้างต้นนี้ส่งผลถึงความสะดวกสบายและใช้งานง่ายของผู้ใช้ เช่น การจัดเรียงลำดับตามราคาสินค้า การค้นหาตามชื่อสินค้าหรือรายละเอียดสินค้า เป็นต้น ดังนั้นการสกัดข้อมูลข้างต้นให้ออกมาอย่างถูกต้องจึงมีส่วนสำคัญอย่างยิ่งต่อโปรแกรมค้นหาสินค้า

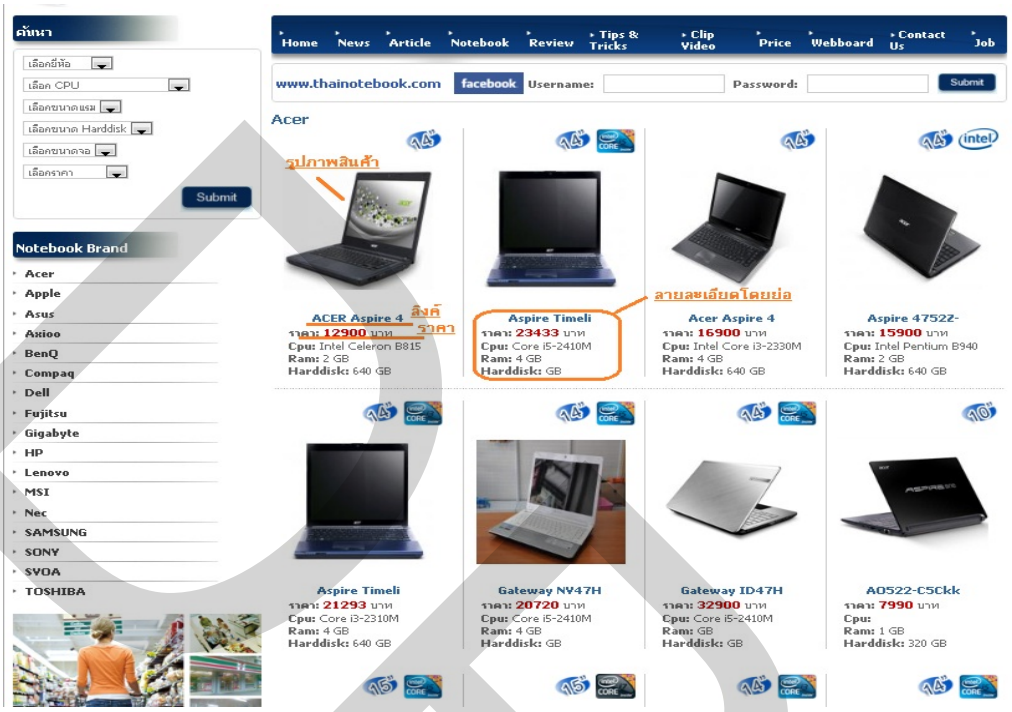


ภาพที่ 3.1 ผลการค้นหาจาก Google Product Search

ที่มา: www.google.com/shopping, 2554, ธันวาคม 20.

3.1.2 ศึกษาเกี่ยวกับเว็บขายสินค้า

จากการศึกษาเว็บขายสินค้าในประเทศไทย ส่วนใหญ่จะมีหน้ารวมสินค้าที่มีการรวมสินค้าหลายๆอย่างที่มีในเว็บไว้ในหน้าเดียว ดังภาพที่ 3.2 โดยจะแสดงข้อมูลแต่ละสินค้าเช่น ราคาสินค้า รูปสินค้า รายละเอียดย่อ และยูอาร์แอลสินค้า ซึ่งจะเห็นว่าเป็นข้อมูลที่ครบถ้วนและจำเป็นสำหรับโปรแกรมค้นหาสินค้าดังที่กล่าวในหัวข้อ 3.1.1 และทำให้เว็บครอเลอร์ที่ใช้งานในโปรแกรมค้นหาสินค้าไม่จำเป็นต้องเข้าไปเก็บทุกๆ ยูอาร์แอลในเว็บขายสินค้านั้นๆ ด้วย ดังนั้นจากประโยชน์ของหน้ารวมสินค้าทำให้เราหาทางสร้างขั้นตอนวิธีการสกัดข้อมูลสินค้าจากหน้ารวมเหล่านี้

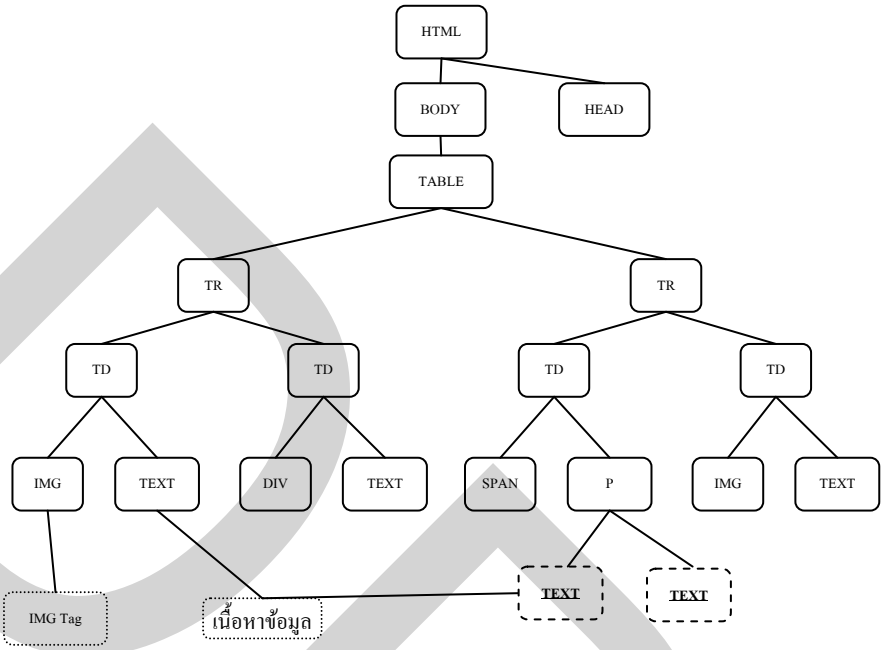


ภาพที่ 3.2 ตัวอย่างหน้ารวมสินค้าจากเว็บขายสินค้า

ที่มา: จาก http://www.computer4u.com, 2554, ธันวาคม 25.

3.1.3 ศึกษาเกี่ยวเว็บครอเลอร์ที่ใช้ในการสกัดข้อมูลสินค้า

จากงานวิจัยที่กล่าวถึงในหัวข้อที่ 2.2 จะพบว่า วิธีการสกัดข้อมูลสินค้านั้นทำได้หลายอย่างตั้งแต่อดีตจนถึงปัจจุบัน จนพบว่าทางเป็นไปได้ที่จะทำการสกัดข้อมูลสินค้าจากเว็บในประเทศไทยนั้น ไม่สามารถใช้เทคโนโลยีเว็บเซอร์วิสได้ ดังนั้นจึงต้องใช้การสกัดข้อมูลจากโครงสร้างภาษา HTML ที่เรียกว่า DOM แทน ซึ่งคือ การสร้างโหนดในของ HTML Tag ขึ้นมา โดยมีโครงสร้างลักษณะต้นไม้ หรือ DOM Tree ดังภาพที่ 3.3 ซึ่งมีลำดับชั้นของโหนดลดหลั่นลงไป โดยโหนดที่อยู่เหนือกว่าเรียกว่า โหนดแม่ โหนดที่อยู่ต่ำกว่าเรียกว่า โหนดลูก และโหนดข้างเคียง คือ โหนดพี่น้อง ซึ่งโหนดแม่จะมีโหนดลูกได้หลายโหนด แต่โหนดลูกสามารถมีโหนดแม่ได้เพียงโหนดเดียว ตัวอย่างเช่น โหนด TEXT ที่ลำดับล่างสุด (ในกรอบเส้นประ) มีโหนดแม่ คือ P ซึ่งจะเห็นว่าโหนด TEXT ทั้ง 2 อันจะมีโหนดแม่ได้เพียงโหนดเดียว แต่โหนด P สามารถมีโหนดลูกเท่าไหร่ก็ได้



ภาพที่ 3.3 ตัวอย่าง DOM Tree

3.2 การออกแบบขั้นตอนและวิธีการสกัดข้อมูลสินค้า

การออกแบบขั้นตอนและวิธีการสกัดข้อมูลสินค้า จะใช้การวิเคราะห์ DOM Tree เพื่อหาโหนดและข้อมูลสินค้าซึ่งแบ่งเป็นขั้นตอน ดังต่อไปนี้

3.2.1 วิเคราะห์ DOM ในหน้ารวมสินค้าของเว็บขายสินค้า

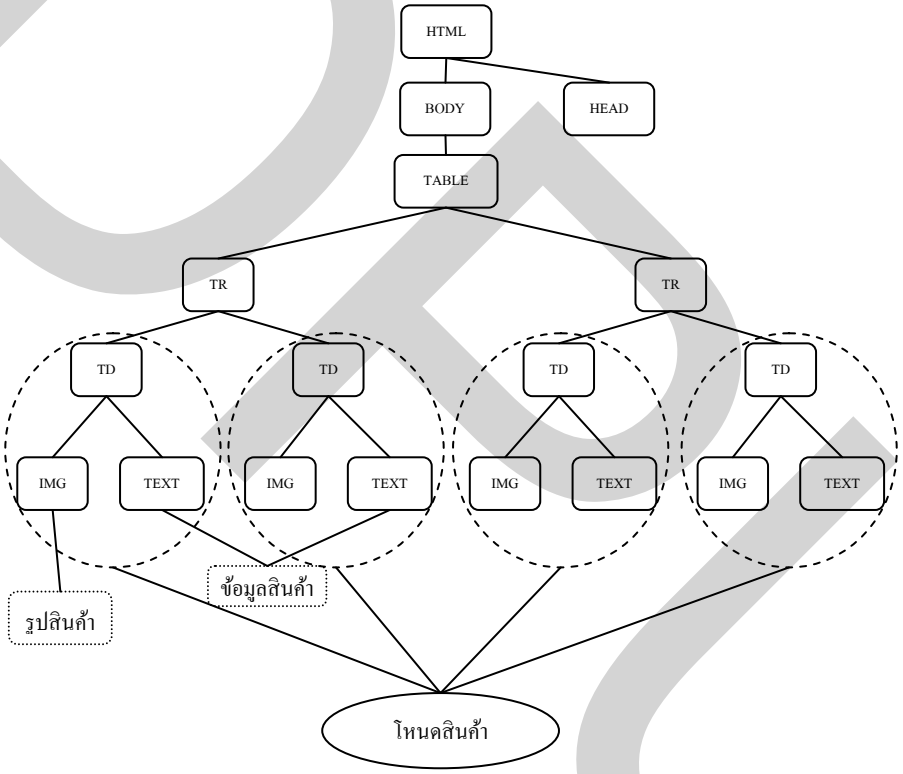
จากการศึกษา DOM ในหน้ารวมสินค้าพบว่า แต่ละสินค้าจะอยู่แยกกันดังภาพที่ 3.4 ดังนั้น ต้องหาวิธีการแยกโหนดแต่ละอันออกจากกัน โดยเรียกโหนดแต่ละอันว่า โหนดสินค้า และสิ่งที่เห็นได้ชัดเจนที่สุดคือ รูปภาพของสินค้า (IMG) ซึ่งจะมี 1 โหนด ต่อ 1 รูปสินค้าเท่านั้น เพราะราคาสินค้าในแต่ละโหนดบางครั้งอาจจะมีหลายราคาเช่น ราคาปกติ ราคาลด ราคาขาย เป็นต้น จึงทำให้เรามุ่งความสนใจไปที่รูปภาพสินค้าว่าควรจะใช้เป็นตัวแยกโหนดสินค้าออกจากกัน แต่ในหน้าเว็บนั้นประกอบไปด้วยรูปภาพต่างๆ ทั้งรูปที่ไม่ใช่รูปสินค้า เช่น ไอคอนและแบนเนอร์ ดังนั้นเราจึงต้องกำจัดรูปที่ไม่จำเป็นเหล่านี้ออกไป

จากการศึกษางานวิจัย (Worasit Choochaiwattana, Winyu Niranatlamphong, and Michael B.Spring., 2007) ซึ่งนำเสนอวิธีการแยกรูปภาพสำคัญในหน้าเว็บ โดยการตัดรูปไอคอนและแบนเนอร์ออกไป ด้วยวิธีการใช้ขนาดของรูปภาพได้แก่ ความกว้าง ความยาว และอัตราส่วน

ภาพ ซึ่งพบว่าเป็นวิธีที่ง่ายและมีประสิทธิภาพ เราจึงนำมาประยุกต์ใช้ในงานวิจัยนี้โดยมีการปรับเปลี่ยนเกณฑ์การกำจัดรูป ดังนี้

1. ความกว้างต้องไม่น้อยกว่า 50 พิกเซล
2. ความยาวต้องไม่น้อยกว่า 50 พิกเซล
3. อัตราส่วนกว้างต่อยาว ระหว่าง 0.5 ถึง 2

เกณฑ์ต่างๆ ได้จาการสำรวจหน้ารวมสินค้าจากเว็บขายสินค้า 20 เว็บและใช้ขนาดที่ต่ำที่สุดที่พบ



ภาพที่ 3.4 ตัวอย่างโหนดในภาษา HTML กับโหนดสินค้า 1 โหนด 1 สินค้า

และเมื่อได้รูปภาพที่ผ่านเกณฑ์ข้างต้น ต่อไปก็จะทำการตรวจสอบว่าโหนดที่คาดว่าจะเป็นโหนดสินค้าคือโหนดใด จากการสำรวจข้างต้น พบว่า แต่ละโหนดสินค้าจะมีรูปภาพเพียง 1 รูป ดังนั้น จึงทำการเลื่อนลำดับชั้นของ DOM Tree จนกระทั่งพบรูปภาพที่ผ่านเกณฑ์มากกว่า 1 รูป จึงระบุได้ว่าโหนดลำดับล่างก่อนหน้าที่จะเลื่อนขึ้นมาครั้งล่าสุดอาจจะเป็นโหนดสินค้า ที่ใช้คำว่าอาจจะเป็นโหนดสินค้าเนื่องมาจากต้องผ่านขั้นตอนการวิเคราะห์หาราคาต่อไปเพื่อให้แน่ใจว่าเป็นโหนดสินค้าจริงๆ

3.2.2 การวิเคราะห์หาราคาสินค้า

โดยใช้การวิเคราะห์ทางภาษาเช่น หน่วยของเงิน เงินบาท (ไทย: บาท; ตัวละติน: Baht; สัญลักษณ์: ฿; รหัสสากลตาม ISO 4217: THB) (encyclopedia.) โดยพิจารณาได้ว่า ตัวแรกที่อยู่ก่อนหน้าหน่วยเหล่านี้ คือ ราคาสินค้า หรือจากคำเริ่มต้นเช่น ราคา price ก็จะพบว่าตัวเลขที่อยู่หลังจากคำเหล่านี้คือ ราคาสินค้า

ข้อมูลที่ใช้ในการวิเคราะห์คือ โหนดที่อาจจะเป็นโหนดสินค้าที่ได้จากการวิเคราะห์ในหัวข้อ 3.2.1 โดยกำจัด HTML Tag ออกไป ให้เหลือเพียงข้อมูลที่เป็น TEXT เท่านั้น เมื่อตรวจพบราคาแล้ว จึงมั่นใจได้ว่านี้คือ โหนดสินค้า จากนั้นจึงหาวิธีสกัดข้อมูลที่ต้องการที่เหลือต่อไป

สำหรับกรณีที่พบราคามากกว่า 1 ราคาเช่น ราคาปกติ กับ ราคาขาย ระบบจะสกัดราคาที่ต่ำกว่าออกมา

3.2.3 การวิเคราะห์หายูอาร์แอลสินค้า

จากการสำรวจหน้าเว็บขายสินค้าพบว่า ยูอาร์แอลสินค้าจะครอบรูปของสินค้าอยู่ดังนี้ รูปสินค้าดังนั้นเราจึงทำการสกัดยูอาร์แอลสินค้าซึ่งอยู่ภายใน A Tag และอยู่ใน Attribute href ออกมา

3.2.4 การวิเคราะห์หารายละเอียดย่อของสินค้า

รายละเอียดย่อของสินค้าเราจะใช้ข้อมูลที่เป็น TEXT ทั้งหมดหลังจากกำจัด HTML Tag ออกไปเป็นรายละเอียดย่อของสินค้า

3.3 การกำหนดแบบแผนการวัดประสิทธิผล

การกำหนดแผนการวัดประสิทธิผลแบ่งเป็นขั้นตอนได้ ดังต่อไปนี้

3.3.1 การสร้างเว็บครอเลอร์เพื่อวัดประสิทธิผล

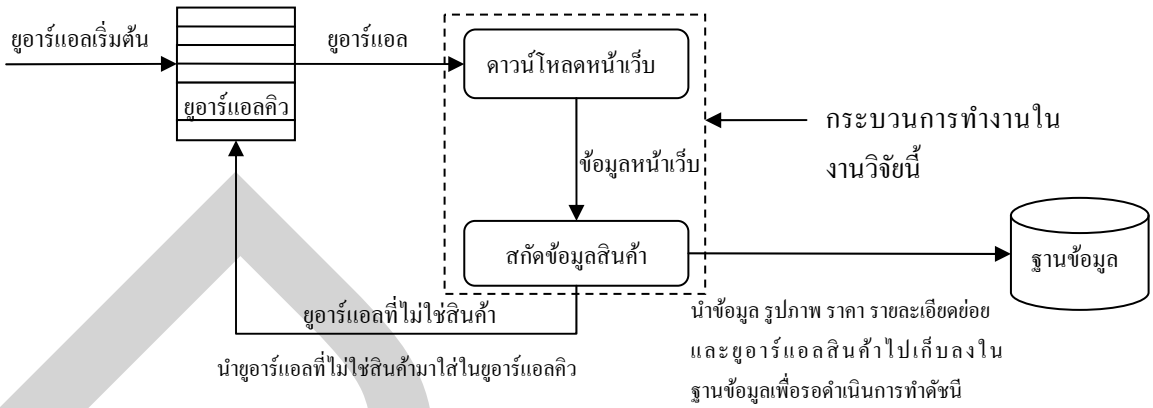
เว็บครอเลอร์โดยรวมมีกระบวนการดังภาพที่ 3.5 ซึ่งสามารถอธิบายได้ดังนี้

1. ยูอาร์แอลเริ่มต้น คือ ยูอาร์แอลแรกที่ต้องการให้เว็บครอเลอร์เข้าไปดำเนินการดาวน์โหลดข้อมูล

2. ยูอาร์แอลคิว คือ ลำดับของยูอาร์แอลที่ต้องการให้เว็บครอเลอร์เข้าไปดาวน์โหลดข้อมูล

3. ฐานข้อมูล คือ ที่เก็บรวบรวมข้อมูลทั้งหมดที่เว็บครอเลอร์ดาวน์โหลดหรือสกัดออกมาได้

4. ส่วนภายในกรอบเส้นประคือส่วนของงานวิจัยนี้



ภาพที่ 3.5 กระบวนการทำงานของเว็บครอเลอร์

ดังนั้นงานวิจัยนี้ไม่ได้สร้างเว็บครอเลอร์แบบสมบูรณ์แต่จะสร้างเฉพาะกระบวนการในกรอบเส้นประเพื่อใช้ในการประเมินประสิทธิผลของขั้นตอนวิธีสกัดข้อมูลสินค้าเท่านั้น ซึ่งเว็บครอเลอร์นี้จะมีหน้าที่ดังนี้

1. ดาวน์โหลดหน้ารวมเว็บขายสินค้า ซึ่งยูอาร์แอลที่ได้นั้นมาจากการคัดเลือกโดยมนุษย์

2. สกัดข้อมูลสินค้าซึ่งจะดำเนินตามขั้นตอนที่ได้จากการวิเคราะห์ในหัวข้อ 3.2

3.3.2 การสร้างหน้าเว็บแสดงผลที่ได้จากการสกัดข้อมูลสินค้าของเว็บครอเลอร์

เนื่องจากขั้นตอนและวิธีการที่เราได้พัฒนานั้นสามารถแยกข้อมูลสินค้าได้ 4 อย่างด้วยกัน คือ รูปสินค้า ราคาสินค้า รายละเอียดย่อสินค้า และยูอาร์แอลสินค้า และใช้การสกัดข้อมูลจากมนุษย์เป็นเกณฑ์ในการวัดประสิทธิผล ดังนั้นจึงต้องออกแบบหน้าแสดงผลให้ง่ายต่อการตรวจสอบความถูกต้องของข้อมูลที่ได้สกัดมา โดยสร้างหน้าเว็บแสดงผลดังภาพที่ 3.6

ภาพที่ 3.6 ตัวอย่างหน้าเว็บผลลัพธ์การสกัดข้อมูลสินค้า

จากภาพที่ 3.6 สามารถอธิบายส่วนต่างๆ ของหน้าเว็บ ตามตัวเลขสีแดงที่ปรากฏ ได้ดังต่อไปนี้

1. ช่องสำหรับกรอกยูอาร์แอลของหน้ารวมสินค้าที่ได้จากการคัดเลือกของมนุษย์
2. แสดงจำนวนผลลัพธ์ที่ได้ทั้งหมด

3. คอลัมน์สำหรับเช็คความถูกต้อง ช่องนี้เป็นช่องสำหรับเช็คข้อมูลความถูกต้องโดยใช้มนุษย์เป็นผู้ตรวจสอบว่าข้อมูลในแถวนั้น มีรายละเอียดใดถูกต้องบ้าง ถ้าถูกต้องก็จะทำการเช็คที่ช่องหน้าข้อมูลแต่ละอัน เช่น image คือ รูปภาพถูกต้อง detail คือ รายละเอียดย่อถูกต้อง link คือ ยูอาร์แอลสินค้าถูกต้อง และ price คือ ราคาสินค้าถูกต้อง

4. คอลัมน์รูปภาพสินค้า ช่องนี้จะแสดงรูปภาพสินค้าที่ได้สกัดออกมา

5. คอลัมน์รายละเอียดย่อกับยูอาร์แอลสินค้า

6. คอลัมน์ปุ่มสำหรับกดเพื่อเช็คว่ายูอาร์แอลที่ได้นั้นตรงกับยูอาร์แอลในหน้าเว็บจริงหรือไม่

7. คอลัมน์ราคาสินค้า

8. ช่องสำหรับกรอกจำนวนสินค้าจากหน้าเว็บจริง ที่ได้จากการตรวจสอบโดยมนุษย์

9. ปุ่มสำหรับกดเพื่อแสดงผลการประเมิน

3.3.3 การคัดเลือกยูอาร์แอลที่ใช้ในการทดลอง

เว็บไซต์ขายสินค้าที่มีอยู่ในประเทศไทยนั้นมีอยู่ด้วยกัน 5 รูปแบบ (ภาวุธ พงษ์วิทยภานุ, 2551) คือ เว็บประมูลสินค้า (E-Auction) เว็บแคตตาล็อคออนไลน์ (Online Catalog Web Site) เว็บประกาศซื้อ-ขาย (E-Classified) เว็บร้านค้าออนไลน์ (E-Shop Web Site) และเว็บตลาดกลางอิเล็กทรอนิกส์ (E-Marketplace) ที่กล่าวรายละเอียดในหัวข้อ 2.1.6

ตารางที่ 3.1 ผลการสำรวจรูปแบบเว็บไซต์ในโปรแกรมค้นหาสินค้า จาก 100 ผลการค้นหา

เว็บไซต์	E- Auction	E- Catalog	E-Shop	E- Marketplace	E- Classified
Google.com/Product	0	3	92	5	0
Pricegrabber.com	3	3	89	5	0

จากการศึกษาโปรแกรมค้นหาสินค้าระดับโลกพบว่าให้ผลดังตารางที่ 3.1 พบว่า เว็บที่เหมาะสมกับโปรแกรมค้นหาสินค้ามีเพียง 3 รูปแบบคือ เว็บร้านค้าออนไลน์ (E-Shop Web Site) เว็บแคตตาล็อคออนไลน์ (Online Catalog Web Site)และเว็บตลาดกลางอิเล็กทรอนิกส์ (E-Marketplace)

แม้ว่าในประเทศไทยจะมีสารบัญเว็บที่แยกหมวดหมู่ของเว็บไซต์ในประเทศไทย แต่ยังคงไม่มีเว็บไซต์ใดที่แยกหมวดหมู่ออกมาได้อยากละเอียดถึงรูปแบบเว็บ E-Commerce ดังนั้นการเลือกยูอาร์แอลจึงทำการเลือกจากโปรแกรมค้นหา Google โดยใส่คำค้นหาเป็นรูปแบบเว็บ

E-Commerce ดังตารางที่ 3.1 โดยเลือกผลลัพธ์ในหน้าแรกจำนวน 3 เว็บต่อคำค้นสำหรับเว็บร้านค้าออนไลน์ และ 10 เว็บต่อคำค้นสำหรับเว็บตลาดกลางอิเล็กทรอนิกส์ โดยเกณฑ์การเลือกนั้นจะสุ่มเลือกจากผลการค้นหา

ตารางที่ 3.2 คำค้นหาที่ใช้ในการสุ่มเลือกจากผลลัพธ์การค้นหาใน Google

ชนิดเว็บไซต์ E-Commerce	**คำค้นหา**	**จำนวนเว็บที่ทำการสุ่มจากผลลัพธ์การค้นหาต่อคำค้น**
เว็บร้านค้าออนไลน์	“ขายหนัง” “ขายเสื้อผ้า” “ขายเครื่องสำอาง” “ขายจีพีเอส” “ขายรองเท้า” “ขายแบตเตอร์รี่” “ขายเครื่องประดับ” “ขายน้ำหอม” “ขายคอม” “ขายเกม”	3
เว็บตลาดกลางอิเล็กทรอนิกส์	“ร้านค้าออนไลน์”	10

คำค้นในตารางที่ 3.2 นั้นได้จากการทดลองหาคำค้นที่ให้ผลลัพธ์เป็นเว็บในรูปแบบที่ต้องการ พบว่า เว็บร้านค้าออนไลน์ใช้คำค้นเป็น “ขาย” กับชนิดสินค้าจะให้ผลลัพธ์เป็นเว็บร้านค้าออนไลน์ แต่ที่แยกเป็นหลายชนิดเพื่อความหลากหลายของข้อมูลที่ใช้ในการประเมินผล สำหรับเว็บชนิดตลาดกลางอิเล็กทรอนิกส์นั้นใช้คำค้นเพียงคำเดียวว่า “ร้านค้าออนไลน์” ก็เพียงพอ เนื่องจากเว็บชนิดตลาดกลางอิเล็กทรอนิกส์นั้น เป็นเสมือนตลาดที่มีสินค้าหลากหลายอยู่แล้ว

จากนั้นเมื่อได้เว็บขายสินค้าที่เลือกมาแล้วจะทำการเลือกหน้าเว็บภายในเว็บไซต์ที่เก็บหน้ารวมสินค้า เช่น หน้าแรก หรือหน้าแยกหมวดหมู่สินค้า เป็นต้น โดย 1 เว็บจะเลือกมา 3 ยูอาร์แอลดังตารางที่ 3.3

ตารางที่ 3.3 ชนิดและจำนวนเว็บไซต์ที่ใช้ในการทดลอง

ชนิดเว็บไซต์ E-Commerce ในไทย	จำนวนหน้าเว็บ
เว็บร้านค้าออนไลน์	30
เว็บตลาดกลางอิเล็กทรอนิกส์	30

3.3.4 ข้อมูลผู้ประเมิน

นักศึกษาในระดับปริญญาโทของมหาวิทยาลัยธุรกิจบัณฑิตย์

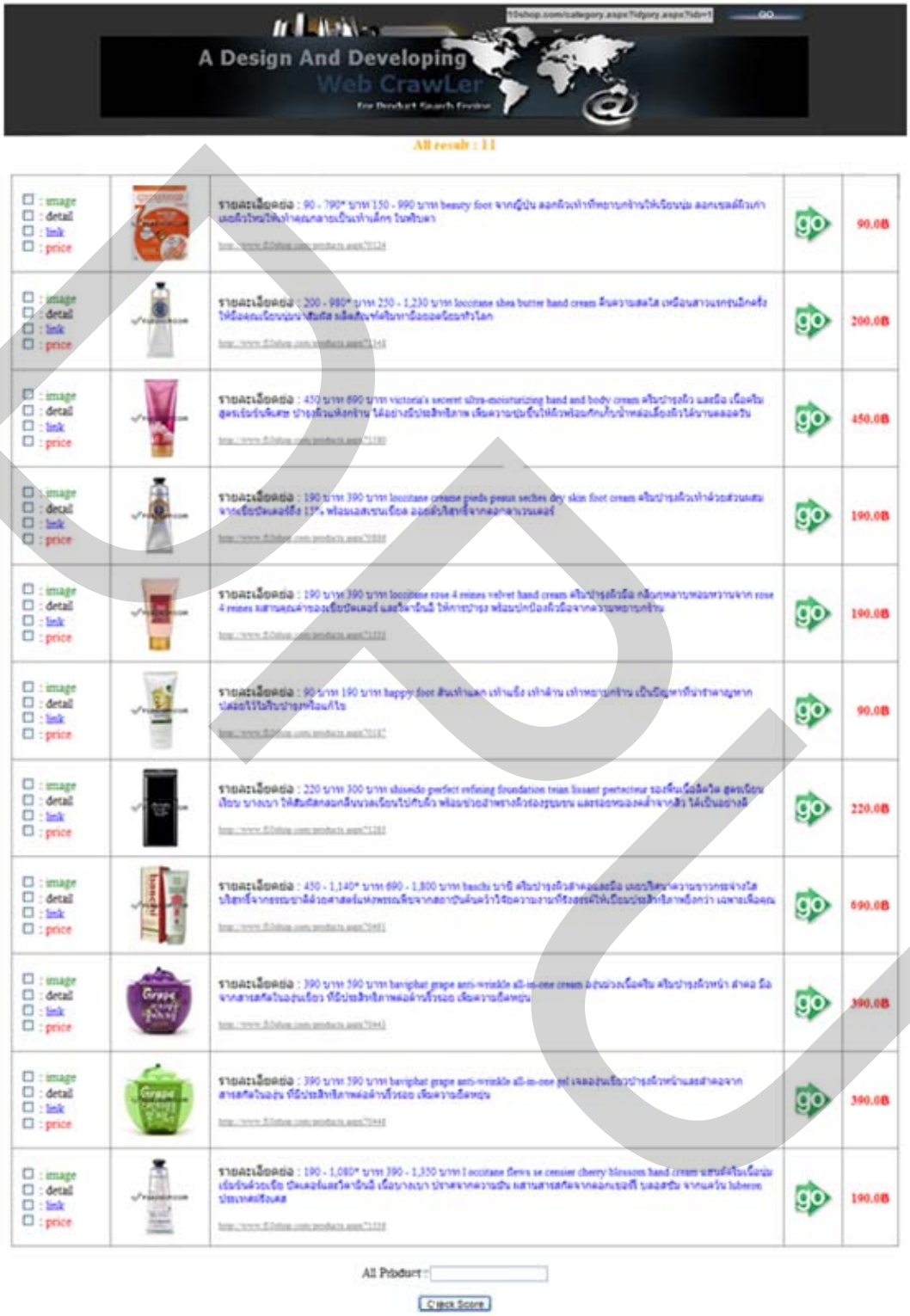
3.4 การวัดประสิทธิผล

จากที่กล่าวไว้ในส่วนของขอบเขตการวิจัยว่า ใช้การสกัดข้อมูลโดยมนุษย์เป็นเกณฑ์ในการวัดประสิทธิผล ดังนั้นขั้นตอนการวัดประสิทธิผล จึงมีขั้นตอนให้ผู้ประเมินปฏิบัติดังนี้

3.4.1 กรอกยูอาร์แอลในช่องหมายเลข 1 ดังภาพที่ 3.7

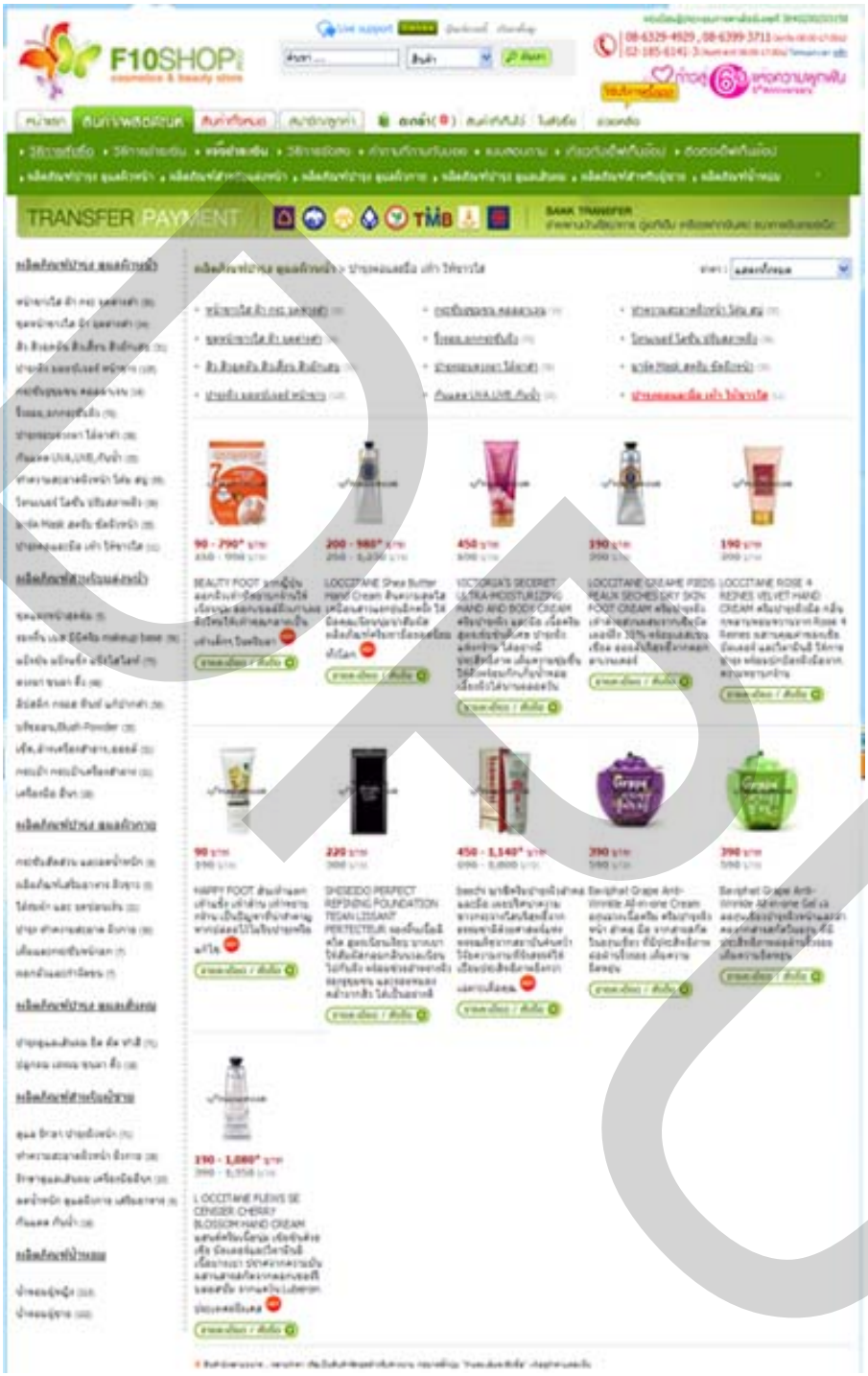
ซึ่งจะขอยกตัวอย่างเป็นยูอาร์แอลต่อไปนี้

http://www.f10shop.com/category.aspx?id=040&pi=0&p=1 โดยการนำยูอาร์แอลดังกล่าวไปกรอกในช่องหมายเลข 1 จากนั้นกดปุ่ม “GO” ระบบจะทำการส่งยูอาร์แอลไปยังเว็บครอเลอร์เพื่อทำการดาว์นโหลด และสกัดข้อมูล เมื่อเว็บครอเลอร์ทำการสกัดข้อมูลเสร็จเรียบร้อย จะส่งการแสดงผลหน้าเว็บ ดังภาพที่ 3.7 ซึ่งผลการสกัดข้อมูลจาก http://www.f10shop.com/category.aspx?id =040&pi=0&p=1 ที่มีหน้าเว็บดังภาพที่ 3.8



ภาพที่ 3.7 ตัวอย่างผลลัพธ์การสกัด 1 สินค้า

ที่มา: http://www.f10shop.com/category.aspx?id =040&pi=0&p=1



ภาพที่ 3.8 ตัวอย่างหน้ารวมสินค้าที่ใช้ในการประเมิน

ที่มา : http://www.f10shop.com/category.aspx?id=040&pi=0&p=1, 2555, กุมภาพันธ์ 26.

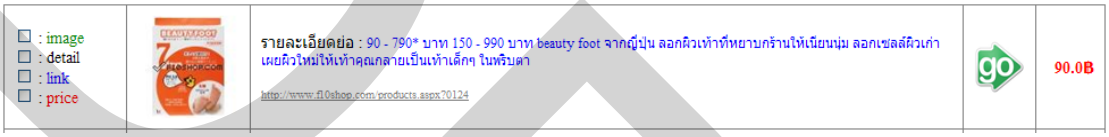
3.4.2 ทำการเปิดหน้าเว็บด้วยยูอาร์แอลเดียวกับข้อที่ 3.3.1 ด้วยเว็บเบราเซอร์ดังภาพที่ 3.8

เมื่อเปิดหน้าเว็บเสร็จแล้วต่อไป ผู้ประเมินจะทำการตรวจสอบ ในหน้าเว็บนั้น ว่ามีสินค้าทั้งหมดกี่สินค้าและนำไปกรอกใส่ช่องที่ 8

3.4.3 ทำการตรวจสอบผลลัพธ์ที่ได้จากขั้นตอนที่ 3.3.1

วิธีการตรวจสอบทำได้โดยการตรวจเช็ค ว่ามีรายละเอียดใดถูกต้องบ้าง ถ้าถูกต้องก็จะทำการเช็คที่ช่องหน้าข้อมูลแต่ละอัน เช่น image คือ รูปภาพถูกต้อง detail คือ รายละเอียดย่อถูกต้อง link คือ ยูอาร์แอลสินค้าถูกต้อง และ price คือ ราคาสินค้าถูกต้อง

การตรวจเช็คนั้นจะใช้ตัวอย่างในภาพที่ 3.9 และ ภาพที่ 3.10 ในการอธิบาย



ภาพที่ 3.9 ตัวอย่างผลลัพธ์การสกัด 1 สินค้า

ภาพที่ 3.10 ตัวอย่างสินค้า 1 สินค้าจากหน้าเว็บ

จากภาพที่ 3.9 และ 3.10 เป็นสินค้าชิ้นเดียวกัน ซึ่งภาพที่ 3.9 เป็นภาพที่ได้จากผลการสกัดข้อมูลสินค้า และภาพที่ 3.10 เป็นภาพจริงจากหน้าเว็บและสามารถตรวจสอบได้ดังนี้

1. รูปภาพเป็นรูปเดียวกันถือว่าการสกัดรูปออกมาได้อย่างถูกต้อง ผู้ประเมินจะทำการเช็คในช่องหน้า image

2. ดูรายละเอียดย่อ คือ เนื้อหาทั้งหมดในโหนดสินค้าซึ่งก็คือเนื้อหาทั้งหมดในภาพที่ 3.10 ซึ่งสกัดออกมาได้ถูกต้อง ผู้ประเมินจะทำการเช็คในช่องหน้า detail

3. ผู้ประเมินทำการกดที่ปุ่ม “GO” จะแสดงหน้าเว็บออกมาดังภาพที่ 3.11 พบว่าเป็นหน้าเว็บเดียวกันกับเมื่อผู้ประเมินกดที่ยูอาร์แอลบนรูปภาพสินค้าในภาพที่ 3.10 ถือว่าการสกัดยูอาร์แอลสินค้าออกมาได้อย่างถูกต้อง ผู้ประเมินจะทำการเช็คในช่องหน้า link

ภาพที่ 3.11 หน้าเว็บเมื่อกดที่ปุ่ม “GO” จากภาพที่ 3.9

ที่มา: http://www.f10shop.com/products.aspx?0124, 2555, กุมภาพันธ์ 26.

4. ราคาที่แยกออกมาได้จากทั้งภาพที่ 3.9 คือ 90 บาทซึ่งขั้นตอนสกัดราคาได้ทำการวิเคราะห์เลือกราคาที่ต่ำที่สินค้าในหน้าขายสินค้านั้นออกมา ซึ่งเมื่อเทียบกับภาพที่ 3.10 พบเป็นราคาที่ต่ำที่สุดคือ 90 บาท ถือว่าการราคาออกมาได้อย่างถูกต้อง ผู้ประเมินจะทำการเช็คในช่องหน้า price

จากนั้นจะดำเนินการตรวจสอบจนครบทุกสินค้าที่แสดงในหน้าผลลัพธ์การสกัดข้อมูล

3.4.4 การคำนวณผลการประเมิน

เมื่อผู้ประเมินได้ทำการดำเนินการตรวจสอบจนครบทุกสินค้าในหน้าผลลัพธ์การสกัดข้อมูลแล้ว ผู้ประเมินจะทำการกดปุ่ม “Check Score” ดังภาพที่ 3.7 หมายเลข 9 จะแสดงหน้าต่างดังภาพที่ 3.12 เพื่อนำผลที่ได้ไปบันทึกข้อมูลและทำการคำนวณค่าความถูกต้อง โดยการคำนวณจะใช้สูตรการคำนวณความถูกต้องดังสมการที่ 1

ภาพที่ 3.12 ตัวอย่างผลการคำนวณจำนวนความถูกต้องโดยมนุษย์ แยกแต่ละรายละเอียด

$$Accuracy = TP + TN / (TP + TN + FP + FN) \quad (1)$$

TP (True Positive) คือจำนวนข้อมูลสินค้าทั้งหมดที่สกัดออกมาได้อย่างถูกต้อง

TN (True Negative) คือจำนวนข้อมูลที่ไม่ใช่สินค้าและไม่ได้ทำการสกัดออกมา

FP (False Positive) คือจำนวนข้อมูลที่ไม่ใช่ข้อมูลสินค้าแต่ถูกสกัดออกมา

FN (False Negative) คือจำนวนข้อมูลสินค้าที่ไม่ได้ถูกสกัดออกมา

จากข้อมูลในภาพที่ 3.12 สามารถคำนวณค่าความถูกต้องได้ ดังตารางที่ 3.4

ตารางที่ 3.4 ตัวอย่างตารางบันทึกค่าความถูกต้องจากยูอาร์แอล

http://www.f10shop.com/category.aspx?id=040&pi=0&p=1

ยูอาร์แอลที่ตรวจสอบ	การสกัดด้วยผู้ประเมิน	การสกัดด้วยวิธีในงานวิจัย				
		TP	TN	FP	FN	Accuracy
http://www.f10shop.com/category.aspx?id=040&pi=0&p=1	11	11	0	0	0	1

3.5 เครื่องมือที่ใช้ในการวิจัย

3.5.1 ฮาร์ดแวร์

PC Computer

Processor : Intel® Core™ i5-2400 CPU @ 3.10 GHz 3.30 GHz

Ram : 4.00 GB

System : Windows 7 Ultimate 32 bit

3.5.2 ซอฟต์แวร์

Firefox และ Netbean 7.0

บทที่4
ผลการศึกษา

4.1 ผลการออกแบบขั้นตอนและวิธีการสกัดข้อมูลสินค้า

จากการออกแบบขั้นตอนและวิธีการสกัดข้อมูลสินค้า ทำให้ได้ผลออกมาเป็นขั้นตอนดังต่อไปนี้

4.1.1 ดาวน์โหลดหน้าเว็บ

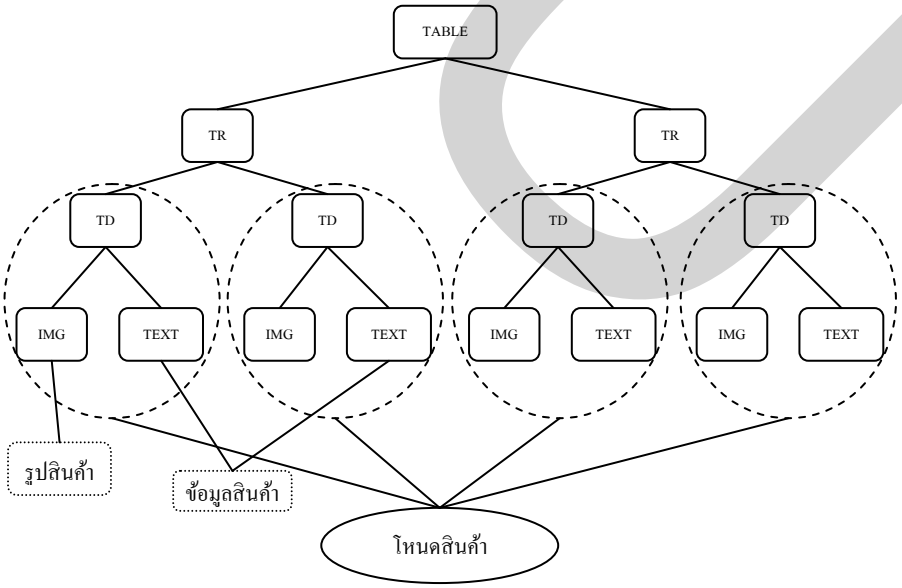
เว็บครอเลอร์ทำการเก็บข้อมูลหน้าเว็บในภาษา HTML

4.1.2 การสกัดโหนดสินค้า

แบ่งออกเป็นกระบวนการย่อยได้ดังรูปที่ 3.3. และมีรายละเอียดดังนี้

4.1.2.1 ทำการสร้างโหนด DOM element ด้วย Htmlparser ไลบรารี่

ทำการสร้างออฟเจ็คในภาษาจาวาที่เป็น โหนดของภาษา HTML ใน DOM Structure โดยมีไลบรารี่สำเร็จรูปในการสร้าง จากนั้นจะสามารถใช้ method ในการจัดการเกี่ยวกับ element ใน DOM Structure ตัวอย่างเช่น getParent(), getChildren() เป็นต้น รวมถึงการคัดกรองชนิดของ HTML Tag เช่น TagNameFilter หรือ AttributeFilter เป็นต้น โดยเราจะใช้ DOM ที่ได้จาก IMG Tag ทั้งหมดในหน้าเว็บนั้นๆ มาดำเนินการ



ภาพที่ 4.1 โครงสร้าง DOM 1 โหนด 1 สินค้า

4.1.2.2 การตรวจสอบรูปภาพสินค้า

จากโครงสร้างเว็บขายสินค้าดังรูปที่ 3.1 แสดงให้เห็นว่าแต่ละสินค้าจะมีส่วนประกอบที่เหมือนกัน และที่เด่นชัดที่สุดคือรูปภาพซึ่งได้มาจากขั้นตอนในหัวข้อ 3.2.3.1 จาก IMG tag ในภาษา HTML ดังนั้นงานวิจัยนี้จึงใช้รูปภาพเป็นปัจจัยแยกในการสกัดข้อมูลสินค้าโดยใช้เกณฑ์ที่ได้จากหัวข้อ 3.2.1 ดังนี้

1. ความกว้างต้องไม่น้อยกว่า 50 พิกเซล
2. ความยาวต้องไม่น้อยกว่า 50 พิกเซล
3. อัตราส่วนระหว่าง 0.5 ถึง 2

4.2.3.3 การเลือกโหนดหรือ Element ที่ใช้ในการสกัดข้อมูลสินค้า

จากตัวอย่างโครงสร้าง DOM ในรูปที่ 4.1 จะเห็นว่าในแต่ละโหนดสินค้าหนึ่งโหนด จะมีรูปสินค้าเพียงรูปเดียวดังโครงสร้างในรูปที่ 3.4 ดังนั้นการเลือกโหนดจึงจะเลือกเฉพาะโหนดที่มีรูปสินค้าเพียงรูปเดียว จากรูปที่ 3.2 จะเห็นว่าจะใช้การคัดกรองเริ่มต้นจาก IMG Tag ดังนั้นจุดเริ่มในแต่ละ Loop จะเข้าไปที่รูปภาพสินค้าโดยตรง แต่จะใช้การตรวจสอบโดยการเลื่อนออกมาที่โหนดแม่ด้วย getParent() เรื่อยๆ จนพบ รูปสินค้า 2 รูปเราก็จะทำการเลือกโหนดก่อนหน้าคือ โหนดล่าสุดที่ยังมีรูปสินค้ารูปเดียวอยู่

4.1.2.4 การสกัดราคาสินค้า

โดยการวิเคราะห์ทางภาษาเช่น หน่วยของเงิน เงินบาท (ไทย: บาท; ตัวละติน: Baht; สัญลักษณ์: ฿; รหัสสากลตาม ISO 4217: THB) โดยพิจารณาได้ว่า ตัวแรกที่อยู่ก่อนหน้าหน่วยเหล่านี้ คือ ราคาสินค้า หรือจากคำเริ่มต้นเช่น ราคา price ก็จะพบว่า ตัวเลขที่อยู่หลังจากคำเหล่านี้คือ ราคาสินค้า

และจากผลการสกัดราคาสินค้าถ้าโหนดใดไม่สามารถสกัดราคาสินค้าออกมาได้ก็จะไม่ถือว่าเป็นโหนดสินค้า และจะไปดำเนินการรับรูปภาพใหม่ตั้งแต่ขั้นตอนที่ 4.2.2.2 อีกครั้งจนกระทั่งไม่มีรูปภาพเหลือในหน้าเว็บนั้น

4.1.3 การสกัดข้อมูลสินค้า

หลังจากที่ได้ราคามาแล้วทำให้เรามั่นใจได้ว่าโหนดนั้นเป็นโหนดของสินค้า ก็จะมาดำเนินการสกัดข้อมูลสินค้าต่อไป

4.1.3.1 สกัดยูอาร์แอลของสินค้า

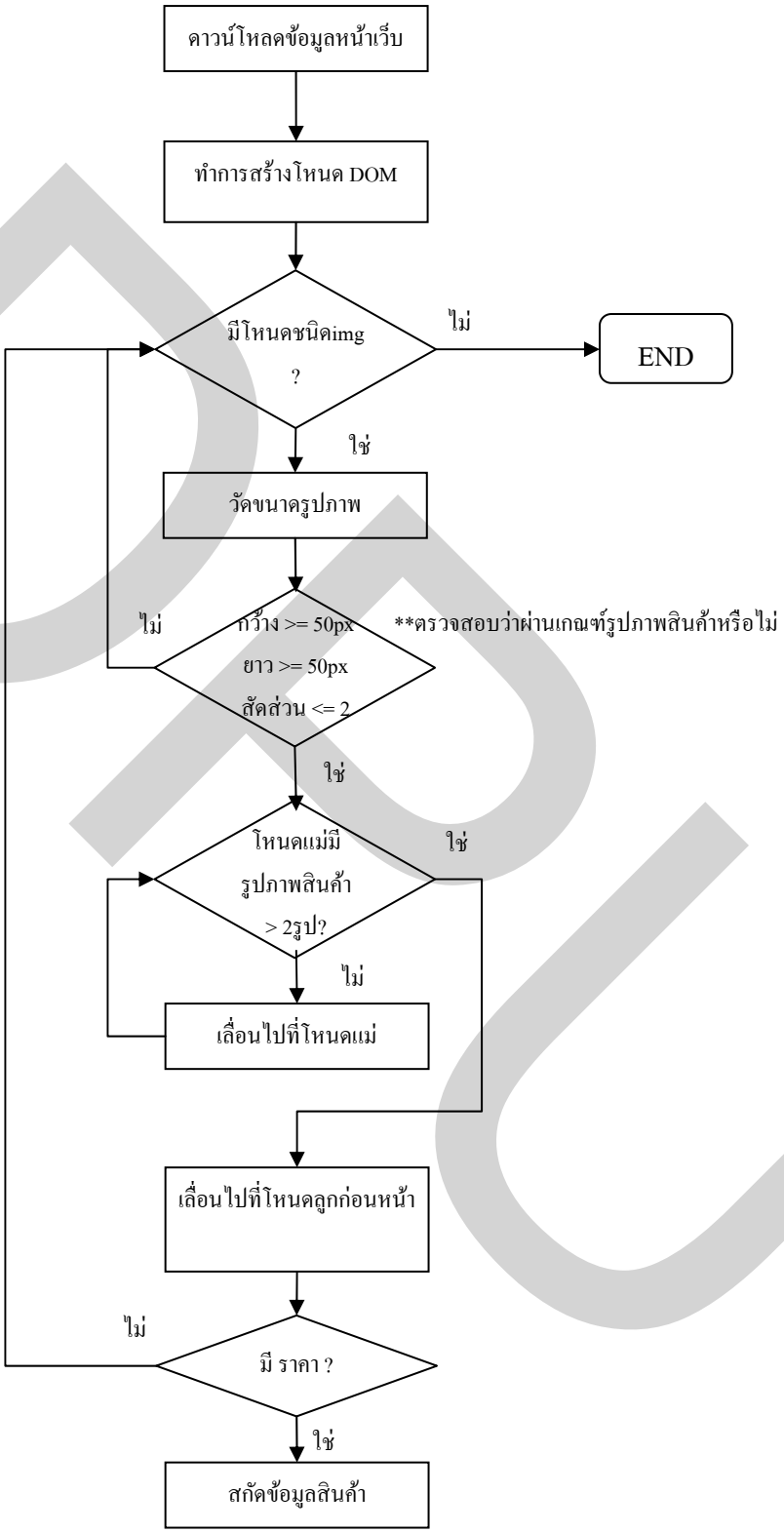
โดยทำการเก็บยูอาร์แอลที่ครอบอยู่กับรูปภาพสินค้าดังรูปแบบภาษา HTML ดังนี้

<imgsrc=”รูปภาพสินค้า” >

4.1.3.2 สกัดรายละเอียดย่อยของสินค้า

โดยทำการเก็บข้อมูลจากโหนดของสินค้าที่ได้ผ่านการตัด HTMLTag ออกไปหมดแล้ว ถึงแม้ว่าในส่วนนี้จะไม่ใช่ชื่อสินค้าโดยตรง แต่จะทำให้ส่วนของการทำดัชนีประหยัดเวลาไปอย่างมากในการหาชื่อของสินค้า เพราะเนื้อหาที่ได้มีเพียงไม่มีก็ตัวอักษรเท่านั้น ต่างกับการทำดัชนีปกติที่ต้องอาศัยการวิเคราะห์จากหน้าเว็บขายสินค้าทั้งหน้าก็คือหน้าเว็บที่เก็บได้จากยูอาร์แอลในข้อ 4.1.3.1

จากขั้นตอนข้างต้นแสดงได้ดัง Flow Chart ดังภาพที่ 4.2



ภาพที่ 4.2 กระบวนการสกัดข้อมูลสินค้า

4.2 ผลการประเมินผลขั้นตอนและวิธีการสกัดข้อมูลสินค้า

ตารางที่ 3.1 ตารางผลการประเมินขั้นตอนการสกัดข้อมูลสินค้า

ชนิดเว็บไซต์ E-Commerce	Accuracy
เว็บร้านค้าออนไลน์และ เว็บไซต์แคตตาล็อกสินค้าออนไลน์	88.4%
เว็บตลาดกลางอิเล็กทรอนิกส์	77.3%

จากตารางที่ 4.1 เป็นผลการประเมินขั้นตอนและวิธีการสกัดข้อมูลสินค้าบนเว็บเพจสำหรับเว็บครอเลอร์ที่ใช้ในโปรแกรมค้นหาสินค้า ซึ่งข้อมูลการประเมินแบ่งได้เป็น 2 ชุดการทดลองคือแถวแรกของหมวดหมู่เว็บร้านค้าออนไลน์และ เว็บไซต์แคตตาล็อกสินค้าออนไลน์ แถวที่สองของหมวดหมู่เว็บตลาดกลางอิเล็กทรอนิกส์ จากตารางจะเห็นว่าค่าความถูกต้องของเว็บตลาดกลางอิเล็กทรอนิกส์ มีค่าที่ต่ำเนื่องจากเว็บรูปแบบนี้มีความซับซ้อนของโครงสร้างหน้าเว็บที่ยุ่งยากมากกว่ารูปแบบแรก

บทที่5
บทสรุป อภิปรายผล และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยชิ้นนี้ได้นำเสนอขั้นตอนและวิธีการสกัดข้อมูลสินค้าสำหรับเว็บครอเลอร์ที่ใช้ในโปรแกรมค้นหาสินค้า โดยการเขียนโปรแกรมย่อยตามขั้นตอนวิธีที่ได้นำเสนอ เพื่อทำการตรวจสอบความถูกต้อง โดยใช้มนุษย์เป็นผู้ตรวจสอบ

จากวัตถุประสงค์ของงานวิจัยจึงขอสรุปเป็นหัวข้อย่อยดังนี้

5.1.1 การนำเสนอแนวทางการเลือกหน้าสินค้า

5.1.2 การออกแบบขั้นตอนและวิธีการสกัดข้อมูลสินค้า

5.1.3 การศึกษาประสิทธิภาพของขั้นตอนและวิธีการสกัดข้อมูลสินค้า

5.1.1 การออกแบบขั้นตอนและวิธีการสกัดข้อมูลสินค้า

จากขั้นตอนและวิธีการสกัดข้อมูลสินค้าที่ได้นำเสนอไปนั้นได้ใช้เกณฑ์วัดเป็นขนาดของรูปภาพเป็นปัจจัยแรก ดังนั้นถ้าเกณฑ์ที่กำหนดผิดพลาดก็ทำให้ไม่สามารถหาโหนดสินค้าเจอได้ แต่จากการวิจัยแสดงให้เห็นแล้วว่าเกณฑ์ที่กำหนดนั้นเป็นที่น่าพอใจ

ดังนั้นถ้าต้องการเพิ่มประสิทธิภาพให้กับขั้นตอนที่ได้นำเสนอนั้น สามารถจัดการกับเกณฑ์ต่างๆ ที่กำหนดทั้ง ขนาดรูปภาพ อัตราส่วนภาพ หรือเกณฑ์การวิเคราะห์ราคาให้เหมาะสมยิ่งขึ้น

5.1.2 การศึกษาประสิทธิภาพของขั้นตอนและวิธีการสกัดข้อมูลสินค้า

ประสิทธิภาพของงานวิจัยนี้จะคำนึงถึงความถูกต้องของข้อมูลสินค้าที่ได้สกัดออกมาโดยใช้มนุษย์เป็นเกณฑ์ และไม่คำนึงถึงความผิดพลาดที่เกิดขึ้นโดยมนุษย์ ดังนั้นตัวมนุษย์เองอาจมีความผิดพลาด (Human Error) บ้างและส่งผลต่อผลการวิจัยบางส่วน แต่เป็นการยอมรับได้เพราะในขณะนี้ยังไม่มีขั้นตอนและวิธีการสกัดข้อมูลสินค้าที่น่าเชื่อถือในภาษาไทย จึงทำให้การใช้มนุษย์เป็นเกณฑ์ เป็นสิ่งที่เหมาะสมที่สุด

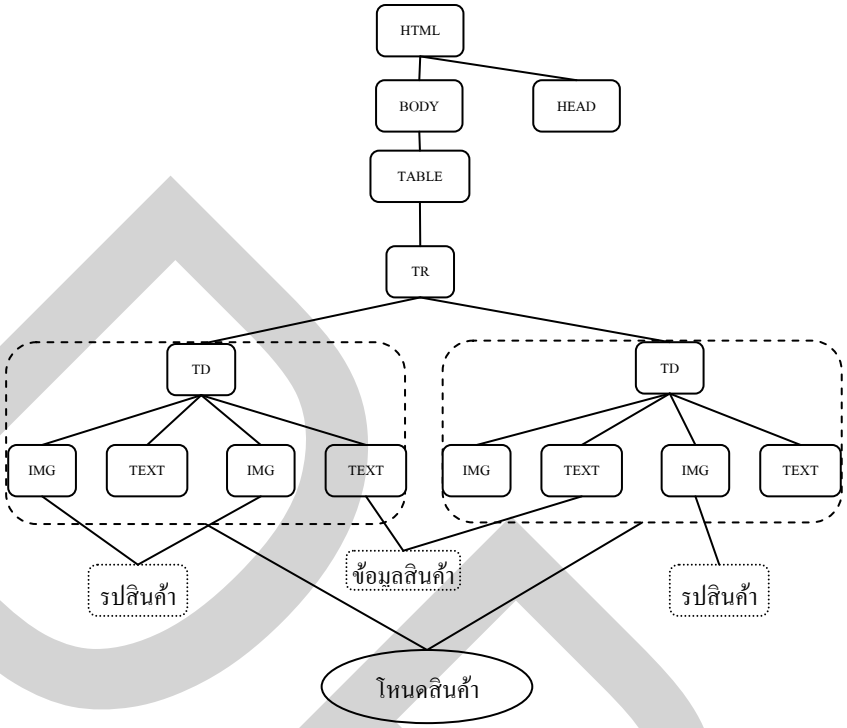
5.2 ปัญหาที่พบในงานวิจัย

ด้วยสมมติฐานที่ว่าสินค้าแต่ละสินค้าอยู่แยกกันในโหนดของภาษา HTML โดยเริ่มดำเนินการวิเคราะห์จากรูปภาพว่าเป็นรูปภาพสินค้าหรือไม่ ถ้าใช่จึงทำการหาโหนดสินค้า จากนั้นวิเคราะห์ต่อว่ามีราคาหรือไม่ ถ้ามีราคาจึงถือว่าเป็นโหนดสินค้า และยังมีเงื่อนไขเพิ่มเติมอีกคือใน 1 โหนดสินค้าต้องมีรูปสินค้าเพียงรูปเดียว ดังนั้นจึงสามารถแบ่งความผิดพลาดของระบบมาวิเคราะห์ได้เป็น 3 ส่วนคือ

5.2.1 ส่วนการสกัดหาโหนดสินค้าผิดพลาด เนื่องจากบางเว็บไซต์ภายในโหนดเดียวมีสินค้ามากกว่า 1 สินค้าหรือรูปสินค้ามากกว่า 1 รูปดังรูปที่ 5.1 ระบบจะไม่สามารถแยกข้อมูลออกมาได้

5.2.2 ส่วนการวิเคราะห์รูปภาพสินค้าผิดพลาด เนื่องจากการตั้งเกณฑ์ ขนาดที่ 50 พิกเซล และอัตราส่วนกว้างยาวไม่เกิน 2 เท่า ส่งผลให้ขาดความยืดหยุ่นในการแสดงผล เนื่องจากบางรูปภาพมีขนาดเล็กกว่า 50 พิกเซล หรือบางสินค้าที่ต้องมีรูปภาพแนวยาวเช่น สร้อยคอหรือชุดกระโปรง ส่งผลให้มีอัตราส่วนมากกว่า 2 เท่า

5.2.3 ส่วนการวิเคราะห์ราคาผิดพลาด เนื่องจากบางเว็บไซต์ไม่มีคำในการช่วยวิเคราะห์หาราคาเช่น ราคา บาท หรือ ฿ แต่มีเพียงชื่อสินค้าและตัวเลขราคา จึงส่งผลให้แม้ว่าเว็บครอเลอร์จะหาโหนดสินค้าเจอแต่ไม่พบราคาสินค้า และบางครั้งส่วนในการวิเคราะห์ราคาก็ยังมีประสิทธิภาพไม่เพียงพอเนื่องจากภายในโหนดสินค้า บางครั้งมีราคามากกว่า 1 ตำแหน่งเช่น ราคาเต็ม ราคาจริง ราคาลด หรือแม้กระทั้งตัวเลขอื่นๆ ที่ใกล้เคียงกับคำช่วยวิเคราะห์หาราคา เพราะบางครั้งหน่วยของราคาสามารถวางได้สองตำแหน่งคือทั้งหน้าและหลังหน่วยเช่น 500 ฿ หรือ ฿ 500 เป็นต้น



ภาพที่ 5.1 ตัวอย่างโหนดในภาษา HTML กับโหนดสินค้า 1 โหนดหลายสินค้า

5.3 ข้อเสนอแนะ

แบ่งตามปัญหาที่พบได้ดังนี้

5.3.1 ส่วนการสกัดหาโหนดสินค้าผิดพลาด ในการสกัดหาโหนดสินค้าควรหาวิธีที่เหมาะสมกับเว็บดังรูปแบบใน รูปที่ 5.1

5.3.2 ส่วนวิเคราะห์รูปภาพสินค้าผิดพลาด ควรมีการวิจัยเพื่อหาเกณฑ์ในการแยกรูปภาพสินค้าที่มีประสิทธิภาพมากยิ่งขึ้น โดยอาจเพิ่มปัจจัยอื่นๆ เช่น เทคโนโลยีการประมวลผลภาพ (Image processing) เข้ามารวมด้วย เช่นการตรวจสอบรูปภาพ ว่าเป็นสินค้าชนิดใด เป็นต้น

5.3.3 ส่วนวิเคราะห์ราคาผิดพลาด

การสร้างเกณฑ์วิเคราะห์ราคาควรมี เกณฑ์การวิเคราะห์ที่นอกเหนือจาก หน่วยสินค้า หรือ คำกำหนดราคา เพราะบางเว็บไซต์ไม่มีทั้งสองอย่าง แต่มีราคาสินค้าอยู่ จึงส่งผลให้ไม่สามารถแยกข้อมูลสินค้าในเว็บไซต์นั้นออกมาได้

ในอนาคตถ้าต้องการนำวิธีการนี้ไปใช้งานจริงในเว็บครอเลอร์ สิ่งที่ต้องคำนึงถึงนอกจากเกณฑ์ต่างๆ ที่กล่าวถึงข้างต้นแล้ว การพัฒนาโปรแกรมหรือรายละเอียดการเขียนโปรแกรมในแต่ละขั้นตอนยังต้องคำนึงถึงเวลาในการประมวลผลและเก็บรวบรวมข้อมูลของเว็บครอเลอร์เป็นสำคัญ เพราะเนื่องจากงานวิจัยนี้ไม่ได้เน้นไปที่โปรแกรม แต่เน้นนำเสนอวิธีการสกัดข้อมูลสินค้าเท่านั้น

บรรณานุกรม

บรรณานุกรม

ภาษาไทย

หนังสือ

ศุภชัย ตั้งวงศ์ศานต์. (2543). **ระบบจัดเก็บและการสืบค้นสารสนเทศด้วยคอมพิวเตอร์ (พิมพ์ครั้งที่ 2).**กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

สารนิพนธ์

นิรันดร์ อังควัฒนวิทย์. (2546). **การเก็บเว็บเพจแบบเฉพาะเจาะจงหัวเรื่องด้วยเว็บคราวเลอร์แบบเรียนรู้ได้.** สารนิพนธ์ปริญญามหาบัณฑิตสาขาวิชาวิศวกรรมคอมพิวเตอร์. กรุงเทพฯ: มหาวิทยาลัยเกษตรศาสตร์

วิริยะแก้วมรินทร์. (2551). **ระบบค้นหาบริการของเว็บเซอร์วิสแบบสื่อความหมายโดยใช้ตัวค้นหาบนเว็บ.** การศึกษาอิสระปริญญาวิทยาศาสตมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น.

สารสนเทศจากสื่ออิเล็กทรอนิกส์

อรวรรณ บึงพรรัตน์. (2549). Search Engine คืออะไร. สืบค้นเมื่อ 12 กันยายน 2554, จาก **http://pirun.ku.ac.th/~g5166319/page/Search%20Engine**

ภาวุธ พงษ์วิทยภานุ. (2551). 5 รูปแบบประเภทของเว็บไซต์ E-Commerce. สืบค้นเมื่อ 20 ธันวาคม 2554, จาก **http://www.pawoot.com/node/327**

ศูนย์วิจัยนวัตกรรมอินเทอร์เน็ตไทย. (2555). สถิติการใช้งานเครื่องมือค้นหาเว็บไซต์. สืบค้นเมื่อ 10 มกราคม 2555, จาก **http://truehits.net/**

encyclopedia. บาท (สกุลเงิน). สืบค้นเมื่อ 20 พฤศจิกายน 2554, from **http://th.wikipedia.org/wiki/บาท_(สกุลเงิน)**

ภาษาต่างประเทศ

BOOKS

B.T. Krulwich, (1996). **The BargainFinder agent: Comparing price shopping on the Internet.**In J. Williams, editor, Bots and other Internet beasties, pp. 258—263. SAMS.NET, Macmillan, 1996.

Roger S. Pressman, David Lowe, (2009). **Web Engineering: A Practioner's Approach 1ED.** McGraw-Hill International Edition, 1996

Roger S. Pressman, David Lowe, (2009). **Web Engineering: A Practioner's Approach 1ED.** McGraw-Hill International Edition, 1996

Doug Cutting, (2005). **Lucene In Action.** Manning Publication Co., 2005

ARTICLES

R.B. Doorenbos, O. Etzioni, and D.S.Weld. (1997). "A scalable comparison-shopping agent for the World-Wide Web ". **In Proceedings of the First International Conference on Autonomous Agents (Agents-97), pp.39—48, 1997.**

Maria Fasli. (2006). "ShopBots: A syntactic present, a semantic futu". **Internet Computing, IEEE, 2006, Vol. 10, Nov-Dec 2006. pp 69-75.Apapanik and Mstefanos. (2007).**

Worasit Choochaiwattana, Winyu Niranatlamphong, and Michael B.Spring. (2007). " Web image classification algorithm: a heuristic rule-based approach s".**In Proceedings of the Second International Conference on Internet Technologies and Application (ITA-07),UK, September 2007**

Yu Chun-Yan, Ma Jun, Zhao Yu-Yan “Online Price Extraction and Decision Support for Agricultural Products,”, **Information Management, Innovation Management and Industrial Engineering, 2009 International Conference, pp 337-340, December 2009**

Robert Baumgartner, Georg Gottlob, Marcus Herzog “Scalable web data extraction for online market intelligence,”, **Proceedings of the VLDB Endowment,Vol 2 Issue 2, August 2009**

ELECTRONIC SOURCES

Encyclopedia. Accuracy and precision. Retrieved January 25, 2012, from **http://en.wikipedia.org/wiki/Accuracy_and_precision**

ภาคผนวก

ภาคผนวก ก
ผลการประเมินขั้นตอนการสกัดข้อมูลสินค้า

ยูอาร์แอลที่ทดสอบ	การสกัดด้วยผู้ประเมิน	การสกัดด้วยวิธีในงานวิจัย				Accuracy
		TP	TN	FP	Fn	
http://www.music-blvd.com/store/	45	45	0	0	0	1
http://www.music-blvd.com/store/product/Electric_Bass-215493-	42	42	0	0	0	1
http://www.music-blvd.com/store/product/Acoustic_Electric_Guitars-213732-th.html	26	26	0	0	0	1
http://www.nbfashion.net/index.php?category=17	16	16	0	0	0	1
http://www.nbfashion.net/index.php?category=15	2	2	0	0	0	1
http://www.nbfashion.net/index.php?modules=nbcute	16	12	0	0	4	0.75
http://www.ready4girl.com/	122	122	0	0	0	1
www.ready4girl.com/เสื้อผ้าแฟชั่นเกาหลี/กระโปรง-skirt	8	8	0	0	0	1
www.ready4girl.com/เครื่องประดับ/กระเป๋าแฟชั่น	10	10	0	0	0	1
http://www.f10shop.com/category.aspx?id=004&pi=0&p=1	30	28	0	0	0	1
http://www.f10shop.com/category.aspx?id=010&pi=0&p=1	30	20	0	0	0	1
http://www.f10shop.com/category.aspx?id=040&pi=0&p=1	10	6	0	0	0	1
http://www.indigoshirt.com/หม้อห้อม-บุรุษ	4	4	0	0	0	1
http://www.indigoshirt.com/index.php?lay=show&ac=cat_showcat&l=1&cid=55413	10	10	0	0	0	1
http://www.indigoshirt.com/index.php?lay=show&ac=cat_showcat&l=1&cid=55845	1	1	0	0	0	1
www.plabin.com	32	29	0	1	2	0.90625
http://www.plabin.com/product_324962_th	3	3	0	0	0	1
http://www.plabin.com/product_398333_th	5	3	0	7	2	0.25
http://www.missblendy.com/	24	24	0	0	0	1
www.missblendy.com/รองเท้าแฟชั่นทุกแบบ/รองเท้าบูทเกาหลี	15	15	0	0	0	1
www.missblendy.com/รองเท้าแฟชั่นทุกแบบ/รองเท้าแตะแฟชั่น	3	3	0	0	0	1
http://www.batt2home.com/	15	15	0	0	0	1
http://www.batt2home.com/index.php?lay=show&ac=cat_showcat&l=1&cid=44471	10	10	0	0	0	1
http://www.batt2home.com/index.php?lay=show&ac=cat_showcat&l=1&cid=49153	10	10	0	0	0	1
http://www.trade-jewelry.com/store/	24	18	0	0	4	0.818182
www.trade-jewelry.com/store/product/แหวน_หลุดจำนำ-	37	36	0	1	0	0.972973
www.trade-jewelry.com/store/product/สินค้าแนะนำ_มาใหม่-1383788-th.html	25	25	0	3	0	0.892857
http://www.orderbyyou.com/s0126/index.php?pgid=cat:24	14	13	0	0	1	0.928571
http://www.orderbyyou.com/s0110/index.php?pgid=	6	0	0	0	6	0
http://www.orderbyyou.com/s0125/index.php?pgid=	20	0	0	0	20	0
เฉลี่ยความถูกต้องในหมวดหมู่เว็บร้านค้าออนไลน์และ เว็บไซต์แคตตาล็อกสินค้าออนไลน์						**0.883961**

ยูอาร์แอลที่ทดสอบ	การสกัดด้วยผู้ประเมิน	การสกัดด้วยวิธีในงานวิจัย				Accuracy
		TP	TN	FP	Fn	
http://www.bangkoksync.com/	25	0	0	0	25	0
www.bangkoksync.com/คอมพิวเตอร์/ส่วนประกอบคอมพิวเตอร์/	1	0	0	0	1	0
www.bangkoksync.com/นาฬิกา-แว่นตา/นาฬิกาข้อมือ/	5	4	0	0	1	0.8
www.tarad.com/	8	8	0	2	0	0.8
Www.tarad.com/category/5994/เครื่องเล่นเกมส์ และเกมส์/Microsoft Xbox 360	34	33	0	0	1	0.970588
http://www.tarad.com/event/topshop?l-id=topshop_060112	35	32	0	0	3	0.914286
http://www.market2u.com/	21	18	0	2	3	0.782609
http://www.market2u.com/mproduct.php?cat=4_1	7	2	0	2	5	0.222222
http://www.market2u.com/mproduct.php?cat=4_7	1	0	0	0	1	0
http://www.weloveshopping.com/portal/plaza/product.php?catid=5,	42	38	0	0	4	0.904762
http://www.weloveshopping.com/portal/plaza/product.php?catid=11	70	70	0	0	0	1
http://www.weloveshopping.com/portal/plaza/product.php?catid=11	17	16	0	0	1	0.941176
http://www.fashionmall.in.th/	38	38	0	0	0	1
http://www.fashionmall.in.th/index.php?app=search&cate_id=274	13	13	0	0	0	1
http://www.fashionmall.in.th/index.php?app=search&cate_id=274&brand=HERMES	7	7	0	0	0	1
http://www.shop2thai.com/plaza.php?cat_id=31	23	21	0	0	2	0.913043
http://www.shop2thai.com/plaza.php?cat_id=62	2	2	0	0	0	1
http://www.shop2thai.com/plaza.php?cat_id=23	4	4	0	0	0	1
http://www.yeepou.com/category/%E0%B8%A3%E0%B8%96%E0%B8%A2%E0%B8%99%E0%B8%95%E0%B9%8C-2	23	20	0	0	3	0.869565
http://www.yeepou.com/category/%E0%B9%80%E0%B8%84%E0%B8%A3%E0%B8%B7%E0%B9%88%E0%B8%AD%E0%B8%87%E0%B8%94%E0%B8%99%E0%B8%95%E0%B8%A3%E0%B8%B5-11	19	15	0	0	4	0.789474
http://www.yeepou.com/category/%E0%B8%99%E0%B8%B2%E0%B8%AC%E0%B8%81%E0%B8%B2%E0%B8%88%E0%B8%B4%E0%B8%A7%E0%B9%80%E0%B8%A7%E0%B8%A5%E0%B8%A3%E0%B8%B5%E0%B9%	44	44	0	0	0	1
http://www.oneclickmarket.com/previewgoos.php?cat_id=117	29	29	0	0	0	1
http://www.oneclickmarket.com/previewgoos.php?cat_id=151	27	27	0	0	0	1
http://www.oneclickmarket.com/plaza.php?cat_id=183	29	29	0	0	0	1
http://www.thaionlinemarket.com/subgroup.asp?gid=9&rid=143	4	4	0	0	0	1
http://www.thaionlinemarket.com/	30	3	0	0	27	0.1
http://www.thaionlinemarket.com/subgroup.asp?gid=2&rid=35	5	5	0	0	0	1
http://1000raan.com/index.php?app=search&cate_id=287&page=2	9	9	0	0	0	1
http://1000raan.com/index.php?app=search&cate_id=3	3	3	0	0	0	1
http://1000raan.com/index.php?app=search&cate_id=20	5	2	0	6	3	0.181818
เฉลี่ยความถูกต้องในหมวดเว็บตลาดกลางอิเล็กทรอนิกส์						**0.772985**

ประวัติผู้เขียน

ชื่อ-นามสกุล	นายกลยุทธ บพิตร
ประวัติการศึกษา	สำเร็จการศึกษาระดับปริญญาตรี สาขาวิชาวิศวกรรมการบินและอวกาศยาน มหาวิทยาลัยเกษตรศาสตร์ ปีการศึกษา 2551
ตำแหน่งและสถานที่ทำงานปัจจุบัน	นักศึกษาทุนผู้ช่วยสอนสาขาวิชาวิศวกรรมเว็บ