ตัวแบบการประเมินเนื้อหาและการสกัดข้อมูลสินค้าเพื่อการคัดแยกเว็บเพจที่เกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์

A Model of Content Evaluation and Product Information Extraction for Electronic Commerce Web Pages Screening

วรสิทธิ์ ชูชัยวัฒนา
ห้องปฏิบัติการวิจัยทางด้านเสิร์ชเอนจินและระบบสารสนเทศอัจฉริยะ
วิทยาลัยครีเอทีฟดีไซน์ แอนด์ เอ็นเตอร์เทนเมนต์เทคโนโลยี มหาวิทยาลัยธุรกิจบัณฑิตย์
110/1-4 ถนนประชาชื่น เขตหลักสี่ กรุงเทพฯ 10210
E-mail: worasit.cha@dpu.ac.th

บทคัดย่อ

การค้นหาข้อมูลบนอินเทอร์เน็ตเป็นหนึ่งในกิจกรรมประจำวันของผู้ใช้งานอินเทอร์เน็ต เสิร์ชเอนจินจึงกลายเป็นเครื่องมือที่สำคัญที่จะช่วยให้ผู้ใช้งานอินเทอร์เน็ตเข้าถึงข้อมูลต่าง ๆ ที่ต้องการได้อย่างมีประสิทธิผลและประสิทธิภาพ งานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อที่จะนำเสนอตัวแบบการประเมินเนื้อหาและสกัดข้อมูลสินค้าเพื่อการคัดแยกเว็บเพจที่เกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์ โดยตัวแบบที่เสนอจะประกอบไปด้วยส่วนประกอบหลัก 2 ส่วนได้แก่ ตัวแบบการประเมินเนื้อหา และ ตัวแบบการสกัดข้อมูลสินค้า ในการทดลองเพื่อประเมินประสิทธิผลของตัวแบบพบว่า เมื่อใช้ตัวแบบการประเมินเนื้อหาร่วมกับตัวแบบการสกัดข้อมูล ให้ค่าเปอร์เซ็นต์ความแม่นยำ 85.56 ดังนั้นตัวแบบที่นำเสนอในงานวิจัยชิ้นนี้ เป็นแนวทางที่น่าสนใจ สำหรับงานที่เกี่ยวกับการประเมินเนื้อหาและการสกัดข้อมูลสินค้าบนเว็บเพจ โดยตัวแบบดังกล่าวสามารถถูกนำมาพัฒนาเป็นกระบวนการทำงานแบบอัตโนมัติที่ฝังไว้ที่เว็บครอว์เลอร์ ซึ่งจะช่วยปรับปรุงการทำงานของเว็บครอว์เลอร์สำหรับระบบเปรียบเทียบราคาแบบอัตโนมัติของสินค้าบนพาณิชย์อิเล็กทรอนิกส์

Abstract

Searching information on the Internet is one of the daily activities of Internet users. Search engines are now the most important tools that can help the Internet users to effectively and efficiently have access to their desired content. This research aims at proposing a model of content evaluation and product information extraction for electronic commerce web pages screening. The model is composed of two main components, which are a content evaluation component and a product information extraction component. The results of experiment revealed that a combination of the content evaluation component and the product information extraction component had an accuracy of 85.56%. Hence, the proposed model was a

promising approach for the task of content evaluation and product information extraction running on web pages. Moreover, the model could be implemented as an automatic process in a web crawler in order to improve the web crawler performance for automatic price comparison systems in electronic commerce business model.

1. บทนำ

ความเจริญก้าวหน้าทางด้านเทคโนโลยีสารสนเทศและเครือข่ายอินเทอร์เน็ตส่งผลกระทบโดยตรงต่อพฤติกรรมในด้านต่าง ๆ ของคนในสังคม ไม่ว่าจะเป็นเรื่องของการติดต่อสื่อสาร การนำเสนอ การแบ่งปัน และการเข้าถึงข้อมูลต่าง ๆ โดยในปี พ.ศ. 2556 มีปริมาณของเว็บเพจที่สามารถสืบค้นได้ผ่านโปรแกรมค้นหา (Search Engine) มากกว่า 40,000 ล้านเว็บเพจและมีแนวโน้มที่จะเพิ่มปริมาณมากขึ้น [1]

ศูนย์วิจัยนวัตกรรมอินเทอร์เน็ตไทย ได้ทำการเก็บข้อมูลจำนวนผู้ใช้อินเทอร์เน็ตในประเทศไทยตั้งแต่ พ.ศ. 2549 – 2559 พบว่ามีจำนวนผู้ใช้อินเทอร์เน็ตในแต่ละวันเพิ่มขึ้นจาก 1,000,000 คนในปี พ.ศ. 2549 กลายเป็นมากกว่า 15,000,000 คนในปี 2559 และในรายงาน Thailand Internet User Profile 2017 ซึ่งจัดทำโดยสำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์ ผู้ใช้งานอินเทอร์เน็ตในประเทศไทยใช้อินเทอร์เน็ตเพื่อเข้าถึงเครือข่ายสังคม (Social Media) และค้นหาข้อมูล (Searching Information) มากที่สุด คิดเป็นร้อยละ 86.9 และ 86.5 ตามลำดับ ในขณะที่ใช้สำหรับการรับ-ส่งอีเมล ร้อยละ 70.5 ใช้ในการดูทีวีและฟังเพลงทางออนไลน์ ร้อยละ 60.7 และใช้อินเทอร์เน็ตสำหรับการซื้อสินค้าและบริการทางออนไลน์มากถึงร้อยละ 50.8 [2] จากรายงานผลการสำรวจมูลค่าพาณิชย์อิเล็กทรอนิกส์ในประเทศไทย ปี 2560 พบว่ามูลค่าอีคอมเมิร์ซปี 2560 เมื่อเทียบกับปี 2559 มีแนวโน้มการเติบโตอย่างต่อเนื่อง โดยมูลค่าอีคอมเมิร์ซประเภท B2C ของปี 2560 มีมูลค่า 812,612.68 ล้านบาท เพิ่มขึ้น 15.54% [3]

อย่างไรก็ดี ผู้ใช้อินเทอร์เน็ตส่วนใหญ่มักจะใช้โปรแกรมการค้นหา (Search Engine) เพื่อการเข้าถึงข้อมูลสินค้าและบริการต่าง ๆ แต่การค้นหาข้อมูลสินค้าและบริการยังมีข้อจำกัดอยู่ เนื่องจากผลลัพธ์ของการค้นคืนจากโปรแกรมการค้นหานั้น จะแสดงผลลัพธ์ที่ได้จากการวัดความเหมือนระหว่างคำค้นและเนื้อหาของเอกสารเท่านั้น ผู้ใช้ที่ทำการค้นหาข้อมูลที่เกี่ยวข้องกับสินค้าและบริการจำเป็นต้องพิจารณาผลลัพธ์การค้นหาด้วยตัวเอง ดังนั้นการพัฒนาระบบเปรียบเทียบราคาแบบอัตโนมัติ (Automatic Price Comparison System) ของสินค้าบนพาณิชย์อิเล็กทรอนิกส์ ซึ่งเป็นรูปแบบหนึ่งของระบบสารสนเทศสมัยใหม่ จึงเป็นทางเลือกที่ท้าทายและน่าสนใจ อย่างไรก็ดี การออกแบบและพัฒนาระบบเปรียบเทียบราคาสินค้าอัตโนมัติให้มีประสิทธิภาพในการทำงาน จะขึ้นอยู่กับการออกแบบและพัฒนาส่วนประกอบที่ทำหน้าที่ในการสกัดข้อมูลจากออกจากเว็บเพจ หากส่วนประกอบดังกล่าวสามารถทำงานได้ดี จะส่งผลทำให้การออกแบบและพัฒนาระบบทำได้ง่ายมากขึ้น

จากการศึกษารวบรวมผลงานวิจัยที่เกี่ยวข้องกับปัญหาดังกล่าว พบว่าเทคนิคการสกัดข้อมูลได้ถูกนำมาใช้เป็นส่วนหนึ่งของการพัฒนาระบบสารสนเทศสมัยใหม่ด้วยเป้าหมายที่แตกต่างกัน เช่น นำมาสกัดข้อมูลที่เกี่ยวข้องกับรายละเอียดวิชา นำมาสกัดข้อมูลที่เกี่ยวข้องกับสินค้า และนำมาสกัดข้อมูลเกี่ยวกับการท่องเที่ยว เป็นต้น อย่างไรก็ดี ความหลากหลายของการรูปแบบการเขียนเว็บ และโครงสร้างเนื้อหาของเว็บส่งผลให้ ปัญหาของการสกัดข้อมูลมีความท้าทายมากยิ่งขึ้น ในงานวิจัยของ [4], [5], [6] และ [7] ได้นำเอาเทคนิคของการจัดกลุ่ม (Clustering Technique) มาประยุกต์ใช้สำหรับการสกัดข้อมูลจากเอกสาร HTML โดยหลักการที่เสนอนั้นสามารถนำไปใช้กับข้อมูลที่ถูกจัดวางอยู่ในรูปของตาราง นอกจากนั้นแล้ว ในงาน [8] ใช้วิธีการวิเคราะห์โครงสร้างของตาราง เป็นหลักการพื้นฐานในการสกัดข้อมูลออกจากเอกสาร HTML ในขณะที่ [9] ได้สรุปไว้ว่าตารางที่อยู่บนหน้าเว็บนั้นมักจะถูกใช้ในการนำเสนอ และการจัดรูปแบบของหน้าเว็บ เนื่องจากในปัจจุบันเริ่มมีการนำเอา Cascading Style Sheet (CSS) เข้ามาช่วยในการตกแต่งและจัดวางรูปแบบของการนำเสนอข้อมูลในเอกสาร HTML ดังนั้นการใช้ตารางในการจัดวางข้อมูลจึงได้รับความนิยมน้อยลง ดังจะเห็นได้จากงาน [10] ซึ่งเป็นการเสนอขั้นตอนวิธี (Algorithm) ในการสกัดข้อมูลสินค้าและความเห็นที่มีต่อสินค้าโดยพิจารณาโครงสร้างเอกสารประกอบกับการใช้พจนานุกรมคำศัพท์ อย่างไรก็ดี งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์โครงสร้างเอกสาร เพื่อการสกัดข้อมูลยังมีอยู่น้อย โดยเฉพาะงานที่ใช้การวิเคราะห์ Document Object Model (DOM) หรือ โครงสร้างของเว็บเพจเพื่อช่วยงานการสกัดข้อมูลสำหรับระบบพาณิชย์อิเล็กทรอนิกส์

ดังนั้นบทความวิจัยชิ้นนี้จะนำเสนอตัวแบบสำหรับการประเมินเนื้อหาบนเว็บเพจและการสกัดข้อมูลสินค้า โดยใช้แนวคิดของการวิเคราะห์โครงสร้างของ Document Object Model สำหรับการคัดแยกเว็บเพจที่เกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์ โดยผู้วิจัยคาดว่าตัวแบบที่พัฒนาขึ้นจะสามารถนำไปพัฒนาต่อให้เป็นกระบวนการอัตโนมัติที่จะฝังไว้ในเว็บครอว์เลอร์ (Web Crawler) สำหรับการพัฒนาเว็บครอว์เลอร์ที่มีความสามารถในการคัดเลือกข้อมูลตามความต้องการได้ในอนาคต

2. สมมติฐานและขอบเขตของงานวิจัย

2.1 สมมติฐาน

ตัวแบบสำหรับการประเมินเนื้อหาบนเว็บเพจโดยใช้แนวคิดของการวิเคราะห์ตัวแบบของเอกสาร (Document Object Model) ซึ่งสามารถใช้ในคัดแยกเนื้อหาของเว็บเพจว่ามีความเกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์ได้อย่างถูกต้องไม่ต่ำกว่า 85% เมื่อเปรียบเทียบกับการประเมินเนื้อหาและสกัดเนื้อหาโดยผู้เชี่ยวชาญ

2.2 ขอบเขตของงานวิจัย

2.2.1 ตัวแบบที่พัฒนาขึ้นจะใช้วิธีการวิเคราะห์ตัวแบบของเอกสาร (Document Object Model) ร่วมกับการวิเคราะห์เนื้อหาของเอกสาร เพื่อประเมินว่าเนื้อหาในเว็บเพจนั้น เกี่ยวข้องกับข้อมูลสินค้าและบริการหรือไม่ ซึ่งจะเป็นการอำนวยความสะดวกให้กับระบบเปรียบเทียบราคาแบบอัตโนมัติ

ของสินค้าบนพาณิชย์อิเล็กทรอนิกส์ ในการคัดเลือกเนื้อหาเพื่อจัดเก็บไว้ในระบบ

2.2.2 การประเมินประสิทธิผลของตัวแบบ จะทำโดยการเปรียบเทียบผลการประเมินเอกสารของตัวแบบและผลการประเมินเอกสารของผู้เชี่ยวชาญ โดยผู้วิจัยถือว่าผลการประเมินเอกสารของผู้เชี่ยวชาญเป็นคำตอบที่ถูกต้อง

2.2.3 งานวิจัยนี้คำนึงถึงประสิทธิผลของตัวแบบเท่านั้น ไม่มีการประเมินประสิทธิภาพของตัวแบบ เช่น เวลาในการทำงานของตัวแบบที่ใช้ในการประเมินเนื้อหาของเว็บเพจ

3. การออกแบบและพัฒนาตัวแบบ

เนื่องจากหน้าที่หลักของตัวแบบสำหรับงานวิจัยนี้ได้แก่การประเมินเนื้อหาและการสกัดข้อมูลสินค้าบนเว็บเพจ ดังนั้นตัวแบบที่เสนอในงานวิจัยนี้จึงถูกแบ่งการทำงานออกเป็นสองส่วน ได้แก่ ส่วนของการพิจารณาเนื้อหาของเว็บเพจ และส่วนของการสกัดเนื้อหา

ในส่วนของการพิจารณาเนื้อหาเว็บเพจ จะมีหน้าที่หลักในการจัดเตรียมเว็บเพจเพื่อใช้สำหรับการสร้างตัวแบบและสำหรับการทำการทดลองวัดประสิทธิผลของตัวแบบ รวมทั้งมีหน้าที่ในการสกัดเนื้อหา เพื่อสร้างคีย์เวิร์ดเวกเตอร์เพื่อเป็นตัวแบบที่ใช้ประเมินเนื้อหาเว็บเพจ ซึ่งส่วนประกอบหลัก (Main Component) ของส่วนการพิจารณาเนื้อหาเว็บเพจได้แก่ 1) เว็บครอว์เลอร์/สไปเดอร์ (Web Crawler/Spider) มีหน้าที่ในการดาวน์โหลดเว็บเพจไว้ที่ 2) คลังเว็บเพจ (Web Page Corpus) ซึ่งเว็บเพจที่ถูกเก็บเอาไว้ที่นี่จะถูกเลือกออกมา (Manual Selected) เฉพาะที่เป็นเว็บเพจที่เกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์ โดยจะถูก 3) ส่วนการสกัดและการเรียนรู้ (Extract Content and Learning) นำเว็บเพจที่เก็บไว้ที่ 4) เทรนนิ่งดาต้าเซ็ต (Training Data Set) ไปสกัดเนื้อหาเพื่อนำมาสร้าง 5) คีย์เวิร์ดเวกเตอร์ (Keywords Vector) โดยขั้นตอนวิธีการสร้าง (Algorithm) แสดงในรูปที่ 1 โดยคีย์เวิร์ดเวกเตอร์จะถูกนำไปใช้ใน 6) ส่วนการระบุเว็บเพจด้านพาณิชย์อิเล็กทรอนิกส์ ซึ่งการทำงานในส่วนนี้ จะนำเอาคีย์เวิร์ดเวกเตอร์ไปเปรียบเทียบกับเว็บเพจที่เก็บไว้ใน 8) เทสติ้งดาต้าเซ็ต (Testing Data Set) ซึ่งถูกเลือกมาด้วยวิธีการสุ่มจากคลังเว็บเพจ เพื่อนำมาพิจารณาว่าเนื้อหาในเว็บเพจดังกล่าวมีความเกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์หรือไม่ โดยจะเก็บผลการประเมินไว้ที่ 8) ผลการประเมินเนื้อหาเว็บเพจ (Candidate Web Pages)

3.1 การสร้างคีย์เวิร์ดเวกเตอร์

กำหนดให้ W เป็นเซตของเว็บเพจที่เว็บครอว์เลอร์ทำการดาวน์โหลดมาเก็บไว้ $W = \{ w_1, w_2, w_3, \ldots, w_n \}$ ในขณะที่ W_{train} เป็นเซตของเว็บเพจที่ถูกเลือกด้วยมือ (Manually Selected Web Pages) จาก W เฉพาะที่มีเนื้อหาเกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์ และ W_{test} เป็นเซตของเว็บเพจที่ถูกเลือกด้วยวิธีสุ่ม (Randomly Selected Web Pages) จาก W ซึ่งจะประกอบไปด้วยเว็บเพจที่มีเนื้อหาเกี่ยวข้องและไม่เกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์ โดยที่ $W_{train} \subseteq W$ และ $W_{test} \subseteq W$ ซึ่งหมายความว่า $\forall x \{ x \in W_{train} \rightarrow x \in W \}$ และ $\forall y \{ y \in W_{test} \rightarrow y \in W \}$ และ $\left| W_{train} \cup W_{test} \right| \leq \left| W \right|$ นอกจากนั้นแล้ว กำหนดให้เว็บเพจ $w_i = (w_i t_j)_{1 \times m}$

$= (w_i t_1, w_i t_2, w_i t_3, \ldots, w_i t_m)$ ซึ่งก็คือเมทริกซ์ขนาด *1 x m* โดยที่ $w_i t$ เป็นจำนวนความถี่ของคำ (Term Frequency) ที่ปรากฏในหน้าเว็บเพจ w_i และ $T_{E\text{-}Commerce} = (t_{1j})_{1 \times p} = (t_{11}, t_{12}, t_{13}, \ldots, t_{1p})$ เป็นเมทริกซ์ขนาด *1 x p* โดยที่ t_{1j} คือความถี่ของคำที่ *j* และ *p* คือจำนวนคำที่รูปแบบคำไม่ซ้ำ (Unique Term Frequency) ในเว็บเพจทั้งหมดที่อยู่ในเซต W_{train}

ดังนั้นในการสร้างคีย์เวิร์ดเวกเตอร์ จะเริ่มจากการนำเอาเว็บเพจใน W_{train} ซึ่งเป็นเว็บเพจที่เกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์ มาทำการสกัดหาเนื้อหาที่บ่งบอกว่าเกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์ โดยใช้อัลกอริทึม *GeneratingKeywordMetic* ดังแสดงในรูปที่ 1 โดยที่ $isNewTerm(T_{E\text{-}Commerce}, w_i, w_i t_j)$ เป็นฟังก์ชันที่ใช้สำหรับการพิจารณาว่าคำดังกล่าวยังไม่มีอยู่ใน $T_{E\text{-}Commerce}$ แต่ในกรณีที่มีคำนั้นอยู่แล้ว $findingTermPosition(T_{E\text{-}Commerce}, w_i, w_i t_j)$ จะถูกเรียกใช้เพื่อหาตำแหน่งของคำดังกล่าวใน $T_{E\text{-}Commerce}$ ในขณะที่ $addNewTermWithTermCount(w_i t_j)$ และ $addCurrentTermWithTermCount(currentTerm, w_i t_j)$ จะเป็นฟังก์ชันที่ทำการเพิ่มความถี่ของคำใหม่ และเพิ่มความถี่ของคำที่ปรากฏอยู่ตามลำดับ นอกจากนั้น *ExtractTopTenTermCount()* เป็นฟังก์ชันที่จะทำสกัดคำที่มีความถี่สูงที่สุด 10 อันดับแรก ออกมาเพื่อใช้เป็นคีย์เวิร์ดเวกเตอร์ในการทำงานในส่วนถัดไป

GeneratingKeyWordCountMetric Algorithm
Input : Set of all Web page W_i in W_{train}
Each Web page W_i that contained all the terms, $w_i t_1$ to $w_i t_m$
Output : Metric T_{Top} that contained unique term top-10 count of all W_i in W_{train}

Step 1 : $T_{E\text{-}Commerce} = (t_{ij})_{1 \times p}$ // initiated $T_{E\text{-}Commerce}$
Step 2 : $T_{Top} = (t_{ij})_{1 \times 10}$ // initiated T_{Top}
Step 3 : **for all** Web Page W_i in W_{train} **do**
  for all $w_i t_j$ with in Web Page W_i **do**
    if $isNewTerm(T_{E\text{-}Commerce}, W_i, w_i t_j)$ **then**
      $T_{E\text{-}Commerce} \leftarrow addNewTermWithTermCount(w_i t_j)$
    else
      $currentTerm = findingTermPosition(T_{E\text{-}Commerce}, W_i, w_i t_j)$
      $T_{E\text{-}Commerce} \leftarrow addCurrentTermWithTermCount(currentTerm, w_i t_j)$
    end if
  end for
end for
Step 4 : $T_{Top} = ExtractTopTenTermCount(T_{E\text{-}Commerce})$
Step 5 : **return** T_{Top}

รูปที่ 1 อัลกอริทึม GeneratingKeywordMetric

3.2 ตัวแบบการประเมินเนื้อหาเว็บเพจ

การประเมินเนื้อหาเว็บเพจว่ามีความเกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์หรือไม่ จะเริ่มจากเว็บเพจในเทสติ้งดาต้าเซ็ต ซึ่งได้แก่ W_{test} จะต้องถูกแปลงเป็นเวกเตอร์ของค่าน้ำหนักของคำในเว็บเพจ เนื่องจากก่อนหน้านี้ได้กำหนดให้เว็บเพจ $w_i = (w_i t_j)_{1 \times m} = (w_i t_1, w_i t_2, w_i t_3, \ldots, w_i t_m)$ ซึ่งก็คือเมทริกซ์ขนาด *1 x m* แต่ $w_i t$ จะเป็นค่าน้ำหนักของคำ (Term Weight) ที่ปรากฏในหน้าเว็บเพจ w_i โดยที่ $w_i \in W_{test}$ เท่านั้น ดังนั้นในการสร้างเวกเตอร์ของค่าน้ำหนักของคำในเว็บเพจจะใช้สมการที่ 1 และกำหนดให้ $term_{ij}$ หมายถึง *term* ที่ *j* ในเว็บเพจ w_i และ n_j เป็นเซ็ตของเว็บเพจ *w* ที่มี $term_j$ ปรากฏอยู่

$$w_i t_j = \frac{|term_{ij}|}{|\sum_{k=1}^{m} term_{ik}|} \cdot \log \frac{|W_{test}|}{|n_j|} \qquad (1)$$

จากนั้นนำ $T_{Top} = (t_{1j})_{1 \times 10} = (t_{11}, t_{12}, t_{13}, \ldots, t_{1j})$ ผลลัพธ์จากอัลกอริทึม GeneratingKeyWordMetric ซึ่งเป็นเมทริซ์ขนาด 1 x 10 เก็บจำนวนความถี่ของคำสูงที่สุด 10 อันดับแรก ที่

สกัดออกมาได้จากเว็บเพจทั้งหมดใน W_{test} โดยจะนำ T_{Top} มาทำการผ่านกระบวนการทำให้เป็นมาตรฐาน (Normalization) ก่อนโดยการใช้สมการที่ 2 โดยกำหนดให้เป็น $T_{Top\text{-}Norm}$

$$t_{1j} = \frac{|t_{1j}|}{\left|\sum_{j=1}^{10} t_{1j}\right|} \quad (2)$$

ตัวแบบการประเมินเนื้อหาเว็บเพจ จะใช้ Cosine Similarity ในการคำนวณคะแนนความเหมือน (Similarity Score) ระหว่าง $T_{Top\text{-}Norm}$ และ w_i ดังสมการที่ 3 และตัวแบบการประเมินเนื้อหาเว็บเพจจะประเมินว่าเว็บเพจนั้นว่ามีความน่าจะเป็นที่จะมีเนื้อหาที่เกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์ โดยพิจารณาว่าคะแนนความเหมือน (Similarity Score) มีค่ามากกว่าค่าขีดแบ่งที่กำหนดไว้ (Predefined Threshold, α) ดังรายละเอียดในสมการที่ 4

$$Sim(w_i, T_{Top-Norm}) = \frac{\sum_{j=1}^{t} w_i t_j \cdot t_{1j}}{\sqrt{\sum_{j=1}^{t} w_i {t_j}^2 \cdot \sum_{j=1}^{t} t_{1j}}} \quad (3)$$

$$Sim(w_i, T_{Top-Norm}) \geq \alpha \quad (4)$$

3.3 วิธีการสกัดเนื้อหา

แม้ว่าตัวแบบการประเมินเนื้อหาเว็บเพจสามารถช่วยกรองเว็บเพจที่มีเนื้อหาที่สามารถนำไปใช้สำหรับการพัฒนาระบบเปรียบเทียบราคาแบบอัตโนมัติได้ในระดับหนึ่ง ดังนั้นเพื่อเพิ่มประสิทธิผลในการทำงาน เว็บเพจที่ผ่านการประเมินโดยมีค่าคะแนนความเหมือน (Similarity Score) มากกว่าค่าขีดแบ่งที่กำหนดไว้ (Predefined Threshold) จะถูกนำมาพิจารณาว่ามีเนื้อหาส่วนใดบ้างสามารถสกัดข้อมูลเกี่ยวกับสินค้าและบริการออกมา เพื่อเตรียมไว้สำหรับการพัฒนาระบบเปรียบเทียบราคาอัตโนมัติต่อไป

ในการพัฒนากระบวนการสกัดข้อมูลจะใช้หลักการวิเคราะห์ตัวแบบวัตถุบนเอกสาร (Document Object Modeling, DOM) โดยกระบวนการดังกล่าวถูกปรับปรุงมาจากงาน [11] โดยตั้งชื่อว่า Extraction Base on Sub-Tree and Sibling Node (ExBSTSN) และกำหนดให้ *CW* เป็นเซ็ตของเว็บเพจ, $CW = \{ cw_1, cw_2, cw_3, \ldots, cw_n \}$ โดยประกอบด้วยเว็บเพจทั้งหมดที่อยู่ในผลการประเมินเนื้อหาเว็บ (Candidate Web Page) โดยที่แต่ละเว็บเพจ cw_m ถูกนำเสนออยู่ในรูปแบบ $nd_1, nd_2, nd_3, \ldots, nd_m$ โดยที่ nd_m เป็นตัวแทนของ Node ใน โครงสร้างต้นไม้ของ DOM (DOM Tree Structure) และกำหนดให้ *E* เป็นเมทริกซ์ขนาด *m x m* แสดงความสัมพันธ์ระหว่าง Node ในโครงสร้างต้นไม้ของ DOM โดยที่ $E(nd_i, nd_j)$ จะเท่ากับ 1 ก็ต่อเมื่อมีเส้นเชื่อม (Edge) ระหว่าง Node nd_i ไปถึง Node nd_j นอกจากนั้นแล้ว กำหนดให้ $NodeType(nd_i)$ และ $ParentNode(nd_i)$ เป็นฟังก์ชันที่จะคืนค่าชนิดของ Node nd_i และเป็นฟังก์ชันที่จะคืนค่าเป็น Node พ่อแม่ (Parent Node) nd_i ตามลำดับ

ในขณะเดียวกันกำหนดให้ $getProductInfoNode(nd_m)$ เป็นฟังก์ชันที่จะคืนค่า Node ลูกของ Node nd_m และ $isProductInfoNode(nd_m)$ เป็นฟังก์ชันที่จะทำการพิจารณาว่า nd_m มีชื่อสินค้าและข้อมูลรายละเอียดสินค้าหรือไม่ และกำหนดให้ $isSiblingNode(nd_i, nd_j)$ และ $get\ SiblingNode(nd_i)$ จะเป็นฟังก์ชันพิจารณาว่า nd_i และ nd_j มี Node พ่อแม่เดียวกัน (Similar Parent Node) หรือไม่ และเป็นฟังก์ชันที่จะคืน Node ที่เป็นพี่น้อง (Sibling Node)

ของ Node nd_i ตามลำดับ และ *ExtractProductInfo*(nd_i, nd_j) เป็นฟังก์ชันที่ทำหน้าที่สกัดเนื้อหาออกจาก Node nd_i และ nd_j รูปที่ 2 แสดงอัลกอริทึม ExBSTSN

4. การทดลองและการประเมินผลตัวแบบ

4.1 การสร้างคลังเว็บเพจ

เพื่อทำการประเมินประสิทธิผลการทำงานของตัวแบบ คลังเว็บเพจ (Web Page Corpus) จึงถูกสร้างขึ้นโดยการส่งเว็บครอว์เลอร์ไปทำการดาวน์โหลดเว็บเพจ โดยที่กำหนดเว็บเริ่มต้น (Seed Web) จำนวนทั้งหมด 56 โดเมนเนม ซึ่งเป็นเว็บที่เกี่ยวข้องและไม่เกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์ และให้เว็บครอว์เลอร์ทำงานในช่วงเดือนกันยายน พ.ศ. 2558 ถึงเดือนมกราคม พ.ศ.2559 ในช่วงระยะเวลาประมาณ 5 เดือน เว็บครอว์เลอร์ได้เข้าไปดาวน์โหลดเว็บเพจต่างๆ จำนวนทั้งสิ้น 35,000 เว็บเพจ และได้ทำการเลือกด้วยวิธีการสุ่ม (Randomly Selected) เว็บเพจในคลังเว็บเพจ เพื่อนำไปไว้ในเทสติ้งดาต้าเซ็ต (Testing Data Set) สำหรับการประเมินการทำงานของตัวแบบ โดยจำนวนเว็บเพจที่อยู่ในเทสติ้งดาต้าเซ็ตมีจำนวนทั้งหมด 900 เว็บเพจ

ExBSTSN Algorithm
Input : Web page W_i that contained all the node n_j to n_k in DOM tree
Metric E that contained edge information between each node in W_i
Output : Extract Result Set, S that contained extracted product information objects

```
Step 1 : S = {}  // initiated S
Step 2 : for all Node nd_i in Web page W_i do
           if NodeType(nd_i) == IMG_TYPE then
             ImgObj ← nil // initiated ImgObj for an image object found in Node nd_i
             ImgObj ← DownloadIMG (nd_i)
             If isProductImage(ImgObj) then
                //Extract Product Info From Sub Tree
                pNode, infoNode ← nil // initiated pNode and infoNode
                pNode ← ParentNode(nd_i)
                if infoNode ← getProductInfoNode(pNode) Then
                   S ← ExtractProductInfoObject(nd_i,infoNode)
                else
                   //Extract Product Info From Sibling Node
                   siblingNode ← nil // initiated siblingNode
                   siblingNode ← getSiblingNode(nd_i)
                   if isProductInfoNode(siblingnode) then
                      S ← ExtractProductInfoObject(nd_i,siblingNode)
                   end if
                end if
             end if
           end if
         end for
Step 3 : return S
```

รูปที่ 2 อัลกอริทึม ExBSTSN

4.2 การออกแบบการทดลองและเมตริกการประเมิน

ในการประเมินประสิทธิผลของตัวแบบการประเมินเนื้อหาและการสกัดข้อมูลสินค้า จะใช้ค่าเปอร์เซ็นต์ความแม่นยำ (Percentage Accuracy) เพื่อวัดความถูกต้องของการประเมินเนื้อหาและการสกัดข้อมูลสินค้า ดังรายละเอียดในสมการที่ 5

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \qquad (5)$$

โดยที่

TP = จำนวนของเว็บเพจที่มีข้อมูลสินค้าที่สกัดออกมาได้อย่างถูกต้อง

TN = จำนวนของเว็บเพจที่ไม่มีข้อมูลสินค้า

FP = จำนวนของเว็บเพจที่สกัดข้อมูลอื่นๆ ออกมาแทนข้อมูลสินค้า

FN= จำนวนของเว็บเพจที่มีข้อมูลสินค้าแต่ไม่สามารถสกัดข้อมูลได้

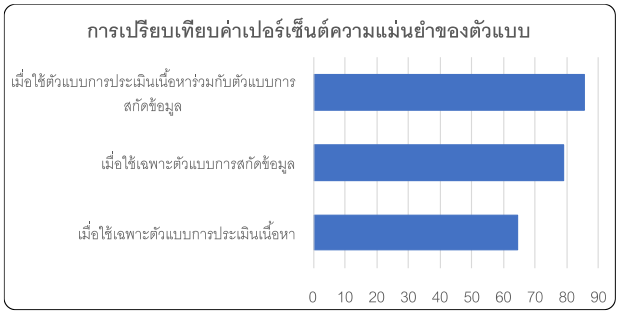
ในการคำนวณค่าความแม่นยำของการทำงานของตัวแบบนั้น ผลลัพธ์ของการทำงานของตัวแบบจะถูกนำไปเปรียบเทียบกับชุดคำตอบที่ถูกต้อง (Golden Standard)

ในการสร้างชุดคำตอบที่ถูกต้อง นักวิจัยและนักศึกษาระดับบัณฑิตศึกษาจากคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์ จำนวนทั้งหมด 3 ท่าน โดยมอบหมายงานให้พิจารณาเนื้อหาเว็บเพจว่ามีข้อมูลสินค้าที่เพียงพอต่อการสกัดข้อมูลออกมาสำหรับระบบเปรียบเทียบราคาอัตโนมัติหรือไม่ เว็บเพจที่อยู่ในเทสติ้งดาต้าเซ็ตจำนวนทั้งหมด 900 เว็บเพจ จะถูกสุ่มขึ้นมาแสดงให้ผู้สร้างชุดคำตอบที่ถูกต้องทีละ 1 เว็บเพจ และให้ผู้สร้างชุดคำตอบพิจารณาประเมิน โดยก่อนที่จะประเมินผู้สร้างชุดคำตอบจะได้รับการแจ้งว่าในการพิจารณานั้น เว็บเพจจะมีประกอบไปด้วย รูปภาพสินค้า ชื่อสินค้า คำอธิบายสินค้า และราคาของสินค้า และเมื่อผู้สร้างชุดคำตอบแต่ละท่าน ทำการประเมินผลเว็บเพจครบถ้วนแล้ว จะนำผลการประเมินของทุกคนมาพิจารณา หากผลการประเมินเว็บเพจเดียวกันของผู้สร้างชุดคำตอบแต่ละท่านมีความขัดแย้งกัน จะดูเสียงข้างมากเป็นหลัก แล้วจึงทำการบันทึกผลการประเมินเว็บเพจดังกล่าวเป็นชุดคำตอบที่ถูกต้อง (Golden Standard) และจะนำชุดคำตอบดังกล่าวไปใช้ สำหรับการคำนวณความแม่นยำของตัวแบบต่อไป

5. ผลการประเมินตัวแบบ

ในการประเมินประสิทธิผลของตัวแบบการประเมินเนื้อหาและการสกัดข้อมูลสินค้าสำหรับระบบเปรียบเทียบราคาอัตโนมัติ จะประเมินประสิทธิผลการทำงานออกเป็น 3 รูปแบบได้แก่ ประสิทธิผลของตัวแบบการประเมินเนื้อหา ประสิทธิผลของตัวแบบการสกัดข้อมูลสินค้า และประสิทธิผลเมื่อใช้ตัวแบบการประเมินเนื้อหาร่วมกับตัวแบบการสกัดข้อมูลสินค้า โดยจะนำเว็บเพจที่อยู่ในเทสติ้งดาต้าเซ็ตจำนวนทั้งหมด 900 เว็บเพจ มาใช้ในการทดลองและทำการเปรียบเทียบผลการทำงานของตัวแบบกับชุดคำตอบที่ถูกต้องที่ได้จากการประเมินโดยผู้เชี่ยวชาญ

รูปที่ 3 แสดงการเปรียบเทียบค่าเปอร์เซ็นต์ความแม่นยำ (Percentage Accuracy) ของตัวแบบการประเมินเนื้อหาและตัวแบบการสกัดข้อมูล



รูปที่ 3 การเปรียบเทียบค่าเปอร์เซ็นต์ความแม่นยำของตัวแบบการประเมินเนื้อหาและตัวแบบการสกัดข้อมูล

จากผลการประเมินความแม่นยำของตัวแบบพบว่า การใช้ตัวแบบการประเมินเนื้อหาร่วมกับตัวแบบการสกัดข้อมูล ได้ผลการทำงานในภาพรวมที่มีความถูกต้องมากที่สุด เมื่อเปรียบเทียบกับการใช้เฉพาะตัวแบบการประเมินเนื้อหา และการใช้เฉพาะตัวแบบการสกัดข้อมูล ดังรายละเอียดในตารางที่ 1 และตารางที่ 2 แสดงข้อมูลการประเมินความแม่นยำและค่าเปอร์เซ็นต์ความแม่นยำ (Percentage Accuracy) ของตัวแบบการประเมินเนื้อหาและการสกัดข้อมูลสินค้าสำหรับระบบเปรียบเทียบราคาอัตโนมัติ ตามลำดับ

ตารางที่ 1 ข้อมูลการประเมินความแม่นยำของตัวแบบสำหรับการประเมินเนื้อหาและการสกัดข้อมูลสินค้า

การประเมินประสิทธิผลการทำงาน	ค่าเปอร์เซ็นต์ความแม่นยำ (Percentage Accuracy)
เมื่อใช้เฉพาะตัวแบบการประเมินเนื้อหา	64.33
เมื่อใช้เฉพาะตัวแบบการสกัดข้อมูล	78.89
เมื่อใช้ตัวแบบการประเมินเนื้อหาร่วมกับตัวแบบการสกัดข้อมูล	85.56

ตารางที่ 2 ค่าเปอร์เซ็นต์ความแม่นยำของตัวแบบสำหรับการประเมินเนื้อหาและการสกัดข้อมูลสินค้า

การประเมินประสิทธิผลการทำงาน	TP	TN	FP	FN
เมื่อใช้เฉพาะตัวแบบการประเมินเนื้อหา	268	311	185	136
เมื่อใช้เฉพาะตัวแบบการสกัดข้อมูล	385	325	66	124
เมื่อใช้ตัวแบบการประเมินเนื้อหาร่วมกับตัวแบบการสกัดข้อมูล	418	352	35	95

เนื่องจากตัวแบบที่นำเสนอในงานวิจัยนี้ จะประกอบไปด้วยตัวแบบการประเมินเนื้อหา และ ตัวแบบการสกัดข้อมูล โดยแต่ละตัวแบบมีวัตถุประสงค์ต่างกัน ดังที่ได้กล่าวมาแล้ว ในหัวข้อที่ 3 การออกแบบและการพัฒนาตัวแบบ เมื่อนำผล การประเมินตัวแบบมาพิจารณาในรายละเอียด จะพบว่า เมื่อใช้เฉพาะตัวแบบการประเมินเนื้อหา จะมีค่าเปอร์เซ็นต์ความแม่นยำเพียงแค่ 64.33% เท่านั้น โดยมีค่า False Positive ค่อนข้างสูง หากพิจารณาวัตถุประสงค์ของการทำงานของตัวแบบประเมินเนื้อหา จะพบว่าในส่วนนี้จะทำการประเมินเว็บเพจเบื้องต้นเพื่อคัดเลือกเป็น Candidate Web Pages ซึ่งหมายถึงเว็บเพจที่มีน่าจะมีเนื้อหาเกี่ยวข้องกับพาณิชย์อิเล็กทรอนิกส์ โดยใช้วิธีการเปรียบเทียบกับคีย์เวิร์ดเวกเตอร์ การเปรียบเทียบในลักษณะนี้ จึงเป็นการกลั่นกรองเบื้องต้นเท่านั้น ในขณะที่การใช้เฉพาะตัวแบบการสกัดข้อมูล จะมีค่าเปอร์เซ็นต์ความแม่นยำสูงกว่า ซึ่งมีสาเหตุมาจากมีค่า False Positive ที่ลดลง ซึ่งเป็นผลมาจากวิธีการสกัดเนื้อหาใน ExBSTSN Algorithm เมื่อพิจารณารายละเอียดของอัลกอริทึมจะพบว่าการใช้โครงสร้างต้นไม้ของ DOM (Document Object Modeling Tree Structure) ช่วยทำให้การวิเคราะห์เนื้อหาทำได้อย่างถูกต้องมากขึ้น อย่างไรก็ดี ค่าของ False Negative ยังคงสูงอยู่ ทั้งนี้ เป็นเพราะลักษณะของอัลกอริทึมมีการกำหนดกฎเกณฑ์ในการประเมินเนื้อหาตามลักษณะการการจัดวางข้อมูล ในบางกรณีกฎเกณฑ์ที่กำหนดไว้ไม่ได้สอดคล้องกับรูปแบบการจัดวางเนื้อหาสินค้าของผู้พัฒนาเว็บเพจ เช่น การใช้โครงสร้าง Table หรือ Division ที่ซับซ้อน ส่งผลทำให้เกิดข้อผิดพลาดในการสกัดเนื้อหาสินค้าด้วยอัลกอริทึมที่เสนอ

ในขณะที่การใช้ตัวแบบการประเมินเนื้อหาร่วมกับตัวแบบการสกัดข้อมูลมีค่าความแม่นยำสูงกว่า ทั้งนี้เป็นเพราะการใช้ตัวแบบทั้งสองร่วมกัน เอื้อต่อการเพิ่มประสิทธิผลในการประเมินเนื้อหาและการสกัดข้อมูลสินค้า เนื่องจากตัวแบบการประเมินเนื้อหาจะมีส่วนช่วยในการคัดกรองเนื้อหาเบื้องต้น ดังนั้นเว็บเพจที่มีเนื้อหาไม่เกี่ยวข้องกับข้อมูลสินค้าจะถูกคัดออกไประดับหนึ่ง และเมื่อนำเว็บเพจที่ผ่านการคัดกรองไปใช้ตัวแบบการสกัดข้อมูล จึงส่งผลทำให้ ลดปริมาณของ False Positive ลง อย่างไรก็ดี การใช้ตัวแบบการประเมินเนื้อหาร่วมกับตัวแบบการสกัดข้อมูล

ยังมีข้อจำกัดอยู่ อันเป็นผลมาจากกฎเกณฑ์ในการประเมินเนื้อหาตามลักษณะการจัดวางข้อมูลตามที่ได้กล่าวมาก่อนหน้านี้ จึงส่งผลให้ False Negative ยังมีปริมาณสูงระดับหนึ่ง

6. บทสรุปและข้อเสนอแนะ

บทความวิจัยชิ้นนี้ นำเสนอตัวแบบสำหรับการประเมินเนื้อหาบนเว็บเพจและการสกัดข้อมูลสินค้า โดยใช้แนวคิดของการวิเคราะห์โครงสร้างของ Document Object Model เพื่อนำไปพัฒนาระบบเปรียบเทียบราคาแบบอัตโนมัติของสินค้าบนพาณิชย์อิเล็กทรอนิกส์ จากผลการประเมินตัวแบบพบว่า การใช้ตัวแบบประเมินเนื้อหาร่วมกับตัวแบบการสกัดข้อมูล ให้ค่าเปอร์เซ็นต์ความแม่นยำ (Percentage Accuracy) เท่ากับ 85.56 ซึ่งสูงกว่าการใช้เฉพาะตัวแบบการประเมินเนื้อหา และการใช้เฉพาะตัวแบบการสกัดข้อมูล ซึ่งให้ค่าเปอร์เซ็นต์ความแม่นยำ เท่ากับ 64.33 และ 78.89 ตามลำดับ

เมื่อพิจารณาผลการประเมินในรายละเอียดพบว่า ตัวแบบที่นำเสนอในบทความวิจัยชิ้นนี้ ยังมีข้อจำกัด และควรจะมีการศึกษาและพัฒนาเพิ่มเติมเพื่อเพิ่มประสิทธิผลในการทำงานของตัวแบบ นอกจากนั้นแล้วการนำเอาตัวแบบดังกล่าวไป พัฒนาต่อให้เป็นกระบวนการอัตโนมัติที่ฝังไว้ในเว็บครอว์เลอร์ (Web Crawler) สำหรับการพัฒนาเว็บครอว์เลอร์ที่มีความสามารถในการคัดเลือกข้อมูลตามความต้องการในอนาคต เป็นสิ่งที่ควรดำเนินการต่อเพื่อที่จะทำการประสิทธิภาพของการทำงานของตัวแบบ เมื่อนำไปฝังไว้ที่เว็บครอว์เลอร์ในอนาคต

กิตติกรรมประกาศ

งานวิจัยชิ้นนี้ ได้รับทุนสนับสนุนการทำวิจัยจากมหาวิทยาลัยธุรกิจบัณฑิตย์ และขอขอบคุณคณาจารย์และนักศึกษาระดับบัณฑิตศึกษาในสาขาวิชาวิศวกรรมเว็บและการพัฒนาแอปพลิเคชันบนอุปกรณ์พกพาที่สละเวลามาเป็นผู้เชี่ยวชาญในการประเมินเนื้อหาเว็บเพจเพื่อสร้างชุดคำตอบที่ถูกต้อง ซึ่งถูกนำมาใช้สำหรับการประเมินตัวแบบที่ได้นำเสนอในบทความวิจัยฉบับนี้

เอกสารอ้างอิง

[1] วรสิทธิ์ ชูชัยวัฒนา, “การปรับปรุงประสิทธิผลของระบบค้นคืนสารสนเทศและโปรแกรมการค้นหา: แนวคิดและเทคนิค”, วารสารวิชาการสมาคมสถาบันอุดมศึกษาเอกชนแห่งประเทศไทย ฉบับวิทยาศาสตร์และเทคโนโลยี ปีที่ 3 ฉบับที่ 1 มกราคม - มิถุนายน พ.ศ. 2557 หน้า 73 - 83

[2] รายงานผลการสำรวจพฤติกรรมผู้ใช้อินเทอร์เน็ตของประ เทศไทย ปี 2560 สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์ (องค์การมหาชน) กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม (https://www.etda.or.th/publishing-detail/thailand-internet-user-profile-2017.html)

[3] รายงานผลสำรวจมูลค่าพาณิชย์อิเล็กทรอนิกส์ในประเทศไทย ปี 2560 สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์ (องค์การมหาชน) กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม (https://www.etda.or.th/publishing-

detail/value-of-e-commerce-survey-2017.html)

[4] Fatima Ashraf, Tansel Özyer, and Reda Alhajj, "Employing Clustering Techniques for Automatic Information Extraction from HTML Document", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) Vol.38 (5) pp. 660-673

[5] Fatima Ashraf and Reda Alhajj, "Clustex : Information Extraction from HTML Pages", Proceeding of 21st International Conference on Advanced Information Networking and Application Workshops (AINAW'07), Vol.1, 2007, pp. 335-360

[6] Kostyantyn Shchekotykhin, Dietmar Jannach, and Gerhard Friedrish, "Clustering Web Documents with Tables for Information Extraction", Proceeding of 4th International Conference on Knowledge Capture (K-CAP'07), 2007, pp. 169-170

[7] Fang Yang, Bo Hu, Cuifen Bai, and Xinyang Han, "Production Information Extraction & Analysis", Proceeding of 3rd International Conference on Communication and Information Processing (ICCIP'17), 2017, pp. 261-267

[8] Paul Bohunsky and Wolfgang Gatterbauer, "Visual Structure-Based Web Page Clustering and Retrieval", Proceeding of 19th International Conference on World Wide Web (WWW'10), 2010, pp. 1067-1068

[9] Mahmoud Shaker, Hamidah Ibrahim, Aida Mustapha, and Lili Nurliyana Abdullah, "Information Extraction from Web Tables", Proceeding of 11th International Conference on Information Integration and Web-based Applications & Services (iiWAS'09), 2009, pp. 470-476

[10] Shenglong Mi, Yinsheng Li, Hao Chen, and Yong Fang, "Extract of Product Information Object for Trustworthiness", Proceeding of 11th IEEE International Conference on e-Business Engineering (ICEBE 2014), 2014, pp. 252-257

[11] Worasit Choochaiwattana "An Algorithm of Product Information Extraction from Web Pages: a Document Object Model Analysis Approach", Proceeding of 2nd International Conference on Information Communication and Management (ICICM 2012), 2012, pp. 103-107