โปรแกรมการดึงข้อมูลจากเว็บไซต์ออกเป็นไฟล์ Excel แบบอัตโนมัติ โดยใช้เทคนิคตามโครงสร้างของเว็บไซต์

ชานนท์ ตั้งสุทธิวงษ์[1] และ อรวรรณ จิตตะกาญจน์[2]
[1]คณะวิทยาศาสตร์และศิลปศาสตร์ สาขาเทคโนโลยีสารสนเทศ มหาวิทยาลัยบูรพา วิทยาเขตจันทบุรี
[2]สาขาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และศิลปศาสตร์ มหาวิทยาลัยบูรพา วิทยาเขตจันทบุรี
Emails: sgraphy-official@hotmail.com, orakik@yahoo.com

บทคัดย่อ

ปัจจุบันข้อมูล ข่าวสารที่ปรากฏอยู่บนเว็บไซต์มีจำนวนมาก และยังเป็นข้อมูลที่ได้อัพเดทอยู่สม่ำเสมอ บทความนี้จึงมีแนวคิดในการจัดทำแอปพลิเคชั่นการดึงข้อมูลจากเว็บไซต์ออกเป็นไฟล์ Excel แบบอัตโนมัติ โดยใช้เทคนิคตามโครงสร้างของเว็บไซต์ โครงงานนี้ได้ทำการเลือกเว็บไซต์วงในซึ่งเป็นเว็บไซต์ที่ให้ข้อมูลเกี่ยวกับร้านอาหาร จะทำการดึงข้อมูล ชื่อร้านค้า ที่อยู่ของร้านค้า เว็บไซต์ เบอร์โทรศัพท์ ละติจูด ลองจิจูด ประเภทอาหาร ชนิดร้านอาหาร วันที่เปิด-ปิด เวลาที่เปิด-ปิด ที่เกี่ยวกับร้านอาหารมาเก็บลงไปยัง Excel ประโยชน์ที่ได้รับจากงานวิจัยนี้คือจะช่วยลด ค่าใช้จ่ายและแรงงาน ในการค้นหา และรวบรวมข้อมูลเกี่ยวกับร้านอาหารทั้งหมด โครงงานนี้พัฒนาด้วยภาษา c# และใช้ Library Html Agility Pack, EPPLUS, Regular Expression

ABSTRACT

Nowadays updated information and news are on several websites. This project is to create an application that automatically pull out website information into excel file based on its structure. Wongnai, restaurants-related website, is pulled out information, name, address, website, telephone number, longitude, latitude, type, open-close date and time, into excel files. The advantages are the decrease in cost of expense and collection of restaurant information. This project is developed by c# and using Library Html Agility Pack, EPPLUS, Regular Expression.

คำสำคัญ—แอปพลิเคชั่น; ข้อมูล; ร้านอาหาร; เว็บไซต์วงใน;

1. บทนำ

ปัจจุบันได้มีเทคโนโลยีทั้งระบบอินเทอร์เน็ต เครือข่ายการสื่อสารต่าง ๆ ได้เข้ามามีบทบาทในชีวิตประจำวัน โดยเฉพาะระบบอินเทอร์เน็ต ซึ่งเป็นเทคโนโลยีที่ช่วยให้ผู้คนรู้ข้อมูลข่าวสารได้ง่าย สะดวก และรวดเร็ว โครงงานนี้จึงมีแนวคิดในการจัดทำแอปพลิเคชั่นการดึงข้อมูลจากเว็บไซต์ออกเป็นไฟล์ Excel แบบอัตโนมัติ โดยใช้เทคนิคตามโครงสร้างของเว็บไซต์ ซึ่งโครงสร้างของเว็บไซต์นั้นไม่มีความแน่นอนหรือตายตัว นักพัฒนาโปรแกรมจึงต้องเขียนตามโครงสร้างของเว็บไซต์นั้นๆ โครงงานนี้ใช้เว็บไซต์วงใน (www.wongnai.com) ซึ่งเป็นเว็บไซต์บริการเครือข่ายสังคมของประเทศไทยที่มีความน่าเชื่อถือ และข้อมูลที่มีการอัพเดทอยู่สม่ำเสมอ ผู้ใช้งานเว็บไซต์สามารถค้นหาร้านอาหาร ข้อมูล รูปและคำวิจารณ์จากสมาชิกคนอื่น สมาชิกผู้ใช้งานเว็บไซต์สามารถเพิ่มข้อมูลร้านอาหารหรือเสนอแนะให้แก้ไขข้อมูลร้านอาหารที่มีอยู่ในฐานข้อมูล แสดงความคิดเห็นเกี่ยวกับร้านอาหาร อัพโหลดรูปภาพเกี่ยวกับร้านอาหาร โดยแอปพลิเคชั่นจะทำการดึงข้อมูล เช่น ชื่อร้านค้า ที่อยู่ของร้านค้า เว็บไซต์ เบอร์โทรศัพท์ ละติจูด ลองจิจูด ประเภทอาหาร ชนิดร้านอาหาร เว็บไซต์ วันที่เปิด-ปิด เวลาที่เปิด-ปิด และข้อมูลอื่น ๆ ที่เกี่ยวกับร้านอาหารมาเก็บลง Excel ซึ่งประโยชน์ที่ได้รับจากโครงงานนี้คือจะช่วยลดทั้งค่าใช้จ่ายและแรงงานในการรวบรวม

ข้อมูลร้านอาหารที่อัพเดทอยู่เสมอ โดยใช้เทคนิคตามโครงสร้างของเว็บไซต์ พัฒนาด้วยภาษา c# และใช้ Library ที่เกี่ยวข้องกับโครงการนี้ เช่น Html Agility Pack, EPPLUS, Regular Expression

2. ทฤษฎีและซอฟต์แวร์ที่เกี่ยวข้อง

2.1 Visual Studio 2015

Visual Studio คือ โปรแกรมตัวหนึ่งที่ช่วยพัฒนาซอฟต์แวร์และระบบต่างๆ เหมาะสำหรับภาษา VB และ VB.NET เนื่องจากไมโครซอฟต์ได้พัฒนาโปรแกรมและภาษาขึ้นมาควบคู่กันเพื่อให้ใช้งานได้ซึ่งกันและกัน ซึ่งนักพัฒนาโปรแกรมจะนำเครื่องมือมาใช้ในการพัฒนาต่อยอดให้เกิดเป็นระบบต่างๆ หรือเป็นเว็บไซต์และแอปพลิเคชั่นต่างๆ

2.2 Mozilla FireFox

Mozilla Firefox คือ โปรแกรมเว็บเบราเซอร์ (Web Browser) ที่ใช้สำหรับเปิดเว็บไซต์โดยมีทีม Mozilla เป็นผู้พัฒนา และยังเป็น (Open source browser) ที่สามารถให้โปรแกรมเมอร์ทั่วโลกพัฒนาโปรแกรมเสริมเพื่อใช้ร่วมกับ Mozilla Firefox ได้อีกด้วย

2.3 Regular Expression Language

มีไว้สำหรับตรวจสอบรูปแบบของ string โดยเราสามารถตรวจสอบรูปแบบใดก็ได้ตามที่ต้องการ เช่น E-mail, Url, Ip address ฯลฯ เพียงแค่กำหนด รูปแบบที่ต้องการตรวจสอบลงไปและยังมีเว็บไซต์ที่คอยช่วย Test Coding ก่อนที่จะนำ pattern นั้นไปใช้จริงในตัว Application

2.4 HtmlAgilityPack

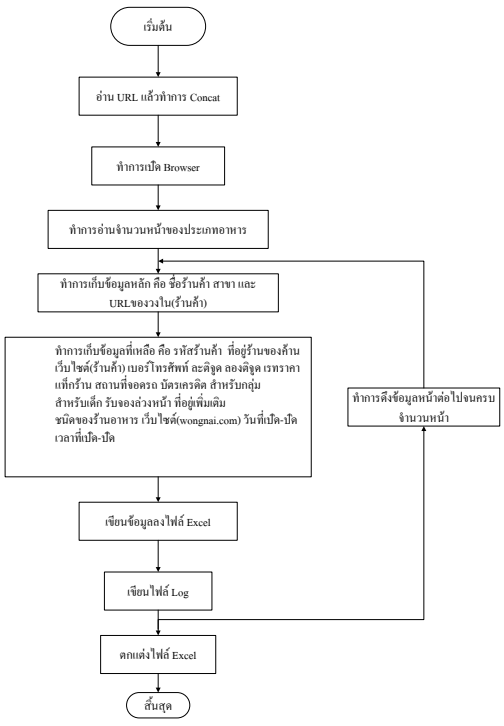
HtmlAgilityPack เป็น Library ที่มีไว้ช่วยในการทำงานกับหน้าเว็บไซต์ ซึ่งสามารถดึงข้อความจากเว็บไซต์ การไต่ Node ลงมาเพื่อดึงข้อมูลจากส่วนนั้น การดึงข้อมูลโดยใช้ Xpath หรือ Innertext ได้

2.5 Selenium - Web Browser Automation (Library in Visual Studio 2015)

ชุดเครื่องมือที่ใช้สำหรับทดสอบเว็บแอปพลิเคชั่นอัตโนมัติ โดยประกอบด้วยเครื่องมือ 4 เครื่องมือ การใช้งานจะขึ้นอยู่กับวัตถุประสงค์ของการทดสอบในแต่ละองค์กร

3. วิธีดำเนินการทดลอง

การจัดทำโปรแกรมดึงข้อมูลจากเว็บไซต์ ในส่วนของผู้ใช้งานโปรแกรมนี้ได้ออกแบบส่วนต่อประสานผู้ใช้ให้ใช้งานง่าย และในส่วนการทำงานของโปรแกรมจะทำงานโดยอัตโนมัติ โดยผู้ใช้งานโปรแกรมสามารถกรอกข้อมูลที่ต้องการจะดึงจากหน้าเว็บไซต์ หลังจากนั้นโปรแกรมจะทำการอ่าน URL และทำการเปิด Browser หลังจากนั้นทางผู้ใช้งานได้ทำการเลือกประเภทอาหารแล้ว โปรแกรมจะทำการเปิดเข้าสู่หน้าประเภทอาหาร ต่อมาทำการอ่านจำนวนหน้าทั้งหมดของประเภทอาหาร แล้วทำการเก็บข้อมูลหลัก ซึ่งหลังจากที่ทำการเก็บข้อมูลหลัก โปรแกรมจะทำการวนลูปเพื่อเข้าเก็บข้อมูลที่เหลือ ซึ่งแต่ละหน้าจะทำการวนลูปจำนวน 20 รอบ เนื่องจากแต่ละหน้ามีร้านอาหารอยู่จำนวน 20 ร้าน และทำการเขียนข้อมูลลง Excel และเขียนไฟล์ Log หลังจากนั้นทำการเก็บข้อมูลในหน้าต่อไปจนครบ และทำการตกแต่งไฟล์ Excel แล้วทำการปิด Browser โดยมีขั้นตอนดังภาพที่ 1

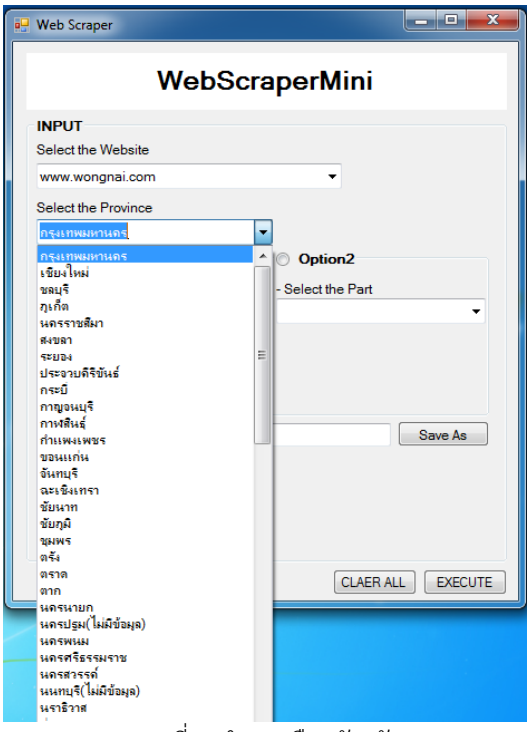


ภาพที่ 1 Flowchart ในส่วนของการ Process

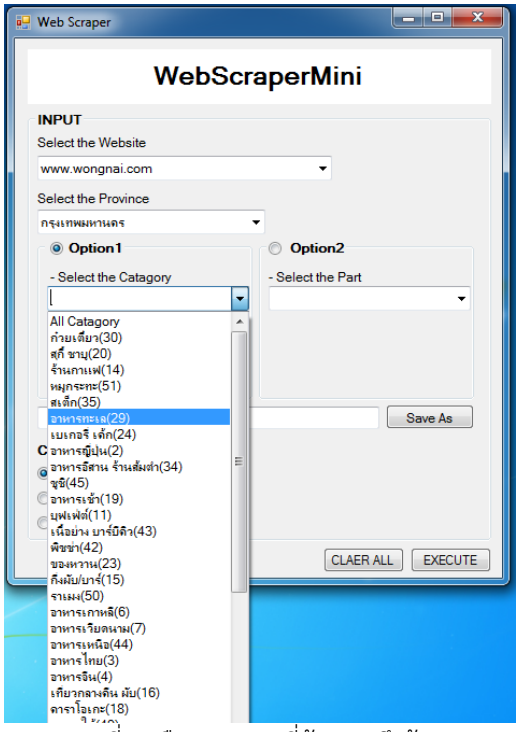
4. ผลการทดลอง

ในการดึงข้อมูลโปรแกรมจะสามารถทำงานได้โดยอัตโนมัติ ผู้ใช้ทำการเลือกข้อมูลเว็บไซต์ ข้อมูลจังหวัด และทางเลือกในการดึงข้อมูล ซึ่งมีอยู่ 3 ทางเลือก คือ ทำการดึงข้อมูลทั้งหมด ทำการดึงข้อมูลเฉพาะประเภทที่เลือก หรือ ดึงข้อมูลเป็นชุด ซึ่ง

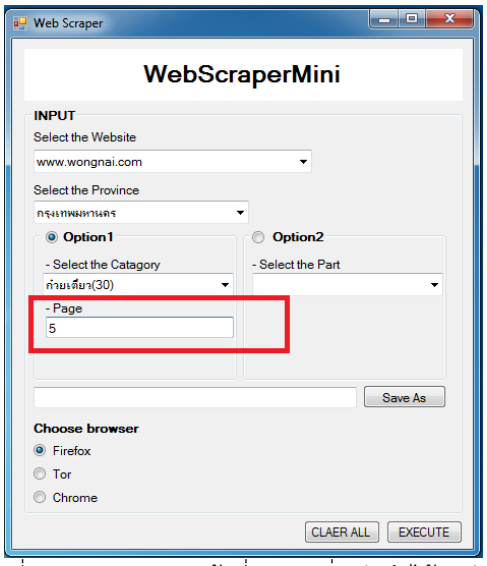
โปรแกรมดึงข้อมูลจากเว็บไซต์ออกเป็นไฟล์ Excel แบบอัตโนมัติ จะทำการดึงข้อมูล รหัสร้านค้า ชื่อร้านค้า ที่อยู่ของร้านค้า เว็บไซต์ เบอร์โทรศัพท์ ละติจูด ลองติจูด เรทราคา แท็กชื่อค้นหาร้าน สถานที่จอดรถ บัตรเครดิต สำหรับกลุ่ม สำหรับเด็ก รับจองล่วงหน้า ที่อยู่เพิ่มเติม สาขา ประเภทอาหาร ชนิดร้านอาหาร เว็บไซต์(wongnai.com) วันที่เปิด-ปิด เวลาที่เปิด-ปิด ดังภาพที่ 2 - 8



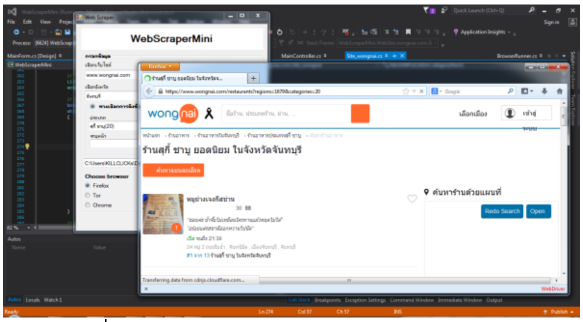
ภาพที่ 2 ทำการเลือกจังหวัด



ภาพที่ 3 เลือกประเภทที่ต้องการดึงข้อมูล



ภาพที่ 4 สามารถกรอกหน้าที่ทำการเริ่มบันทึกได้ หลังจากข้อมูลครบแล้วทำการกด Execute เพื่อทำการดึงข้อมูล



ภาพที่ 5 ตัวอย่างการรันข้อมูลและทำการดึงข้อมูล

	Id	Name	Street_Address	Website
1	Id	Name	Street_Address	Website
2		ก๋วยเตี๋ยวหลุดโลก	ซ.ศิริมังคลาจารย์ 13 ถ.ศิริมังคลาจารย์เชียงใ	https://www.facebook.co
3	130254	ยเตี๋ยวอัญชัน	ถนน ศิริมังคลาจารย์ ซอย 9เชียงใหม่	https://www.facebook.co
4		ก๋วยเตี๋ยวต้มยำไข่หวาน	ถนน อารักษ์เชียงใหม่	https://www.facebook.co
5	146279	กระหรี่หมี่เตี๋ยว	ตรงทางเข้ากองบิน 41 ฝั่งตลาดต้นพยอมเชีย	https://www.facebook.co
6	166826	ซุ้มเฮียฮ่ง	ลานประเสริฐแลนด์ ข้างร้านโจ๊กบ้าหม้อเชียง	https://www.facebook.co
7	23460	สุดยอด@นิมมานฯ	ถนนนิมมานเหมินท์ ตรงข้ามซอย 13เชียงใ	https://www.facebook.co
8	13952	ข้าวซอยนิมมาน	นิมมานเหมินท์ ซอย 7เชียงใหม่	https://www.facebook.co
9	16914	อ่องทิพย์รส (อ่องช้าง	ถนนพระปกเกล้าเชียงใหม่	
10		รสดีเยี่ยมโคขุน	ถนนรัตนโกสินทร์เชียงใหม่	
11	24090	ข้าวซอยแม่สาย	ถนนราชพฤกษ์เชียงใหม่	
12	185901	ก๋วยเตี๋ยวผู้พัน	30/1 ม.7 ต.สันกลาง อ.สันกำแพง จ.เชียงใ	https://www.facebook.co
13	15209	ก๋วยเตี๋ยวลีลา	ถนนหมื่นด้ามพร้าคตเชียงใหม่	https://www.facebook.co
14	181982	ครัวคุณแม่เชียงใหม่	เจ็ดยอด-ช่างเคี่ยนเชียงใหม่	https://www.facebook.co
15	8324	คุณหมอคูซีน	นิมมานเหมินท์ ซอย 17เชียงใหม่	https://www.facebook.co
16	7790	เนื้อตุ๋นรสเยี่ยม	26 ถ.นิมมานเหมินทร์ ซ.11เชียงใหม่	
17	7813	เย็นตาโฟศรีพิงค์	ถนน สุเทพเชียงใหม่	
18	178505	ก๋วยเตี๋ยวเรืออองตอง	ถนน ศิริมังคลาจารย์ ซอย 11เชียงใหม่	https://www.facebook.co
19	8414	ร้านในสวนอารมณ์ดี	สุเทพเชียงใหม่จากสี่แยกไฟแดงตลาดต้นพย	https://www.facebook.co
20	118963	ก๋วยเตี๋ยวโคคา สุกี้ช้า	มณีนพรัตน์เชียงใหม่	https://www.facebook.co
21		นกแล ก๋วยเตี๋ยวต้มยำ	ถนนช้างคลานเชียงใหม่	
22	10473	บะหมี่เกี๊ยวแมน	ถนน เจริญเมืองเชียงใหม่	https://www.facebook.co
23	26168	ก๋วยเตี๋ยวเรือตี๋น้อย	ซอย สุเทพเชียงใหม่ร้านตั้งอยู่ตรงข้ามประตู	
24	158023	ก๋วยเตี๋ยวห้อยขาริมปิง	เชียงใหม่-ลำพูนเชียงใหม่	https://www.facebook.co
25	155122	บะหมี่กังฟูแพนด้า	ถนน สุเทพเชียงใหม่	https://www.facebook.co
26	11980	ร้านข้าวซอยคุณยาย	ศรีภูมิ 8เชียงใหม่	

wongnaiPOI

ภาพที่ 6 ตัวอย่างผลลัพธ์ไฟล์ Excel

PhoneNO	Latitude	Longitude	ResLocate	
088-978-7996, 086-541-6646	18.79337732344101	98.97134599140168	ก๋วยเตี๋ยวstreetfood	ต่ำกว่
084-949-2828	18.79423839198281	98.97243154844477		ต่ำกว่
089-431-4040	18.789597	98.978347	ก๋วยเตี๋ยว#ก๋วยเตี๋ยวต้มยำไข่หว	ต่ำกว่
088-261-1007	18.789604256594696	98.96259967361061	รสจัดจ้านถึงใจกินเอามันส์สังสร	ต่ำกว่
093-935-4869	18.797675356808167	98.96226581186056		ต่ำกว่
08-1648-8238	18.79667788706477	98.96586542518332	อาหารไทยกินเอามันส์อิ่มอร่อย	ต่ำกว่
053-894 881	18.79801200068058	98.96927060061171		ต่ำกว่
053-221-921	18.794536	98.986947		ต่ำกว่
085-030-3993	18.801557666775377	99.01074271576022		ต่ำกว่
053-213-284	18.7995773928679	98.9752097754631	อาหารไทยอาหารเหนือ	ต่ำกว่
094-628-4412, 081-980-6085	18.784675	99.053993	รสจัดจ้านถึงใจกินเอามันส์สังสร	101 -
053-211-765, 086-911-0022	18.80466072056609	98.98676209151745		ต่ำกว่
095-594-7443	18.807751	98.965894		ต่ำกว่
053-226-379	18.795231722546834	98.96679818956659		101 -
081-472-9619	18.796835	98.968373		ต่ำกว่
053-809-129	18.789527	98.965666	อาหารไทยต้นตำรับรสจัดจ้านถึง	ต่ำกว่
053-216-696	18.79385	98.971285		ต่ำกว่
081-764-6883	18.788641785657344	98.95955425396733	อาหารไทยรสจัดจ้านถึงใจอิ่มอร	ต่ำกว่
081-5946597	18.795717294842127	98.9857215889549		ต่ำกว่
084-500-1047,089-701-8057	18.776338193936308	98.99910058826208		ต่ำกว่
053-248-502	18.787017	99.009737		ต่ำกว่
085-811-6335	18.793406803729486	98.95289915211492		ต่ำกว่
053-244405	18.781206418229303	99.00598762946515		ต่ำกว่
081-527-3343	18.792507	98.956259		ต่ำกว่
05-321-1663,08-3208-7092	18.795238	98.983302	ร้านกาแฟบรรยากาศดีอาหารไท	ต่ำกว่

ภาพที่ 7 ตัวอย่างผลลัพธ์ไฟล์ Excel

www.wongnai.com.xlsx_Part 1(1-20)_filelog - Notepad

File Edit Format View Help

```
16/11/2559 16:18:16: Catagory(30) Page(1) = success
16/11/2559 16:19:02: Catagory(30) Page(2) = success
16/11/2559 16:19:57: Catagory(30) Page(3) = success
16/11/2559 16:20:51: Catagory(30) Page(4) = success
16/11/2559 16:21:45: Catagory(30) Page(5) = success
16/11/2559 16:22:37: Catagory(30) Page(6) = success
16/11/2559 16:23:21: Catagory(30) Page(7) = success
16/11/2559 16:24:12: Catagory(30) Page(8) = success
16/11/2559 16:25:01: Catagory(30) Page(9) = success
16/11/2559 16:25:46: Catagory(30) Page(10) = success
16/11/2559 16:26:25: Catagory(30) Page(11) = success
16/11/2559 16:27:04: Catagory(30) Page(12) = success
16/11/2559 16:27:44: Catagory(30) Page(13) = success
16/11/2559 16:28:24: Catagory(30) Page(14) = success
16/11/2559 16:29:03: Catagory(30) Page(15) = success
16/11/2559 16:29:42: Catagory(30) Page(16) = success
16/11/2559 16:30:20: Catagory(30) Page(17) = success
```

ภาพที่ 9 ตัวอย่างไฟล์ Log

5. สรุปผลการทดลองและข้อเสนอแนะ

โครงงานนี้จัดทำแอปพลิเคชั่นการดึงข้อมูลจากเว็บไซต์ออกเป็นไฟล์ Excel แบบอัตโนมัติ โดยใช้เทคนิคตามโครงสร้างของเว็บไซต์ โดยใช้เว็บไซต์วงใน (www.wongnai.com) เป็นซึ่งเป็นเว็บไซต์บริการเครือข่ายสังคมของประเทศไทยที่มีความน่าเชื่อถือและข้อมูลที่มีการอัพเดทอยู่สม่ำเสมอ เป็นกรณีศึกษา โดยแอปพลิเคชั่นจะทำการดึงข้อมูล เช่น ชื่อร้านค้า ที่อยู่ของร้านค้า เว็บไซต์ เบอร์โทรศัพท์ ละติจูด ลองจิจูด ประเภทอาหาร ชนิดร้านอาหาร เว็บไซต์ วันที่เปิด-ปิด เวลาที่เปิด-ปิด และข้อมูลอื่นๆ ที่เกี่ยวกับร้านอาหารมาเก็บลง Excel ซึ่งประโยชน์ที่ได้รับจากโครงงานนี้คือจะช่วยลดทั้งค่าใช้จ่ายและแรงงานในการรวบรวมข้อมูลร้านอาหารที่อัพเดทอยู่เสมอ และประโยชน์ที่นำไปใช้ต่อสามารถนำข้อมูลนี้ไปอัพเดทให้กับหน่วยงานที่ใช้ข้อมูลจะทำให้ได้ข้อมูลปัจจุบัน

ข้อเสนอแนะคือถ้ามีการเปลี่ยนแปลงโครงสร้างของเว็บไซต์จะต้องมีการปรับปรุงการเขียนคำสั่งในโปรแกรม ซึ่งจะพัฒนาเทคนิคที่ใช้ได้กับทุกโครงสร้างต่อไป ซึ่งแนวทางในการพัฒนาต่อ นักพัฒนาโปรแกรมต้องสร้างเงื่อนไขเพื่อรองรับในการเปลี่ยนโครงสร้างของหน้าเว็บไซต์ ซึ่งโครงสร้างของเว็บไซต์สามารถเปลี่ยนแปลงได้ตลอดเวลา ซึ่งจะต้องสร้างเงื่อนไขออกมาเพื่อรองรับให้ใช้ได้ทุกกับโครงสร้าง

เอกสารอ้างอิง

[1] กิตินันท์พลสวัสดิ์ .Visual C# 2010. พิมพ์ครั้งที่ 1 นนทบุรี : สำนักพิมพ์ไอดีซี, 2554.

[2] เกรียงศักดิ์ จันทีนอกการเขียนโปรแกรมทางธุรกิจ .1. พิมพ์ครั้งที่1 : มหาสารคามอภิชาตการพิมพ์, 2551.

[3] Anders Hejlsberg, Scott Wiltamuth, and Peter Golde, The C# Programming Language (Second Edition), Addison-Wesley, 2006.

[4] Ian Graham, Object-Oriented Methods (Third Edition), Addison-Wesley, 2001.

[5] Matt Weisfeld, The Object-Oriented Thought Process (Second Edition), SAMS Publishing, 2003.

[6] Stephen R. Schach, Object-Oriented and Classical Software Engineering (Seventh Edition), McGraw-Hill, 2006.