

โปรแกรมการดึงข้อมูลจากเว็บไซต์ออกเป็นไฟล์ Excel แบบอัตโนมัติ โดยใช้ เทคนิคตามโครงสร้างของเว็บไซต์

ชานนท์ ตั้งสุทธีวงศ์¹ และ อรรณณ จิตตะกาญจน์²

¹คณะวิทยาศาสตร์และศิลปศาสตร์ สาขาเทคโนโลยีสารสนเทศ มหาวิทยาลัยบูรพา วิทยาเขตจันทบุรี

²สาขาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และศิลปศาสตร์ มหาวิทยาลัยบูรพา วิทยาเขตจันทบุรี

Emails: sgraphy-official@hotmail.com, orakik@yahoo.com

บทคัดย่อ

ปัจจุบันข้อมูล ข่าวสารที่ปรากฏอยู่บนเว็บไซต์มีจำนวนมาก และยังเป็นข้อมูลที่ได้อัปเดตอยู่เสมอ บทความนี้จึงมีแนวคิดในการจัดทำแอปพลิเคชันการดึงข้อมูลจากเว็บไซต์ออกเป็นไฟล์ Excel แบบอัตโนมัติ โดยใช้เทคนิคตามโครงสร้างของเว็บไซต์ โครงงานนี้ได้ทำการเลือกเว็บไซต์ซึ่งเป็นเว็บไซต์ที่ให้ข้อมูลเกี่ยวกับร้านอาหาร 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—แอปพลิเคชัน; ข้อมูล; ร้านอาหาร; เว็บไซต์วงใน;

1. บทนำ

ปัจจุบันได้มีเทคโนโลยีทั้งระบบอินเทอร์เน็ต เครือข่ายการสื่อสารต่าง ๆ ได้เข้ามามีบทบาทในชีวิตประจำวัน โดยเฉพาะระบบอินเทอร์เน็ต ซึ่งเป็นเทคโนโลยีที่ช่วยให้ผู้คนรู้ข้อมูลข่าวสารได้ง่าย สะดวก และรวดเร็ว โครงงานนี้จึงมีแนวคิดในการจัดทำแอปพลิเคชันการดึงข้อมูลจากเว็บไซต์ออกเป็นไฟล์ Excel แบบอัตโนมัติ โดยใช้เทคนิคตามโครงสร้างของเว็บไซต์ ซึ่งโครงสร้างของเว็บไซต์นั้นไม่มีความแน่นอนหรือตายตัว นักพัฒนาโปรแกรมจึงต้องเขียนตามโครงสร้างของเว็บไซต์นั้นๆ โครงงานนี้ใช้เว็บไซต์วงใน (www.wongnai.com) ซึ่งเป็นเว็บไซต์บริการเครือข่ายสังคมของประเทศไทยที่มีความน่าเชื่อถือ และข้อมูลที่มีการอัปเดตอยู่เสมอ ผู้ใช้งานเว็บไซต์สามารถค้นหาร้านอาหาร ข้อมูล รูปและคำวิจารณ์จากสมาชิกคนอื่น สมาชิกผู้ใช้งานเว็บไซต์สามารถเพิ่มข้อมูลร้านอาหารหรือเสนอแนะให้แก่ไขข้อมูลร้านอาหารที่มีอยู่ในฐานข้อมูล แสดงความคิดเห็นเกี่ยวกับร้านอาหาร อัปโหลดรูปภาพเกี่ยวกับร้านอาหาร โดยแอปพลิเคชันจะทำการดึงข้อมูล เช่น ชื่อร้านค้า ที่อยู่ของร้านค้า เว็บไซต์ เบอร์โทรศัพท์ ละติจูด ลองจิจูด ประเภทอาหาร ชนิดร้านอาหาร เว็บไซต์ วันที่เปิด-ปิด เวลาที่เปิด-ปิด และข้อมูลอื่นๆ 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 หรือเป็นเว็บไซต์และแอปพลิเคชันต่างๆ

2.2 Mozilla FireFox

Mozilla Firefox คือ โปรแกรมเว็บเบราว์เซอร์ (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

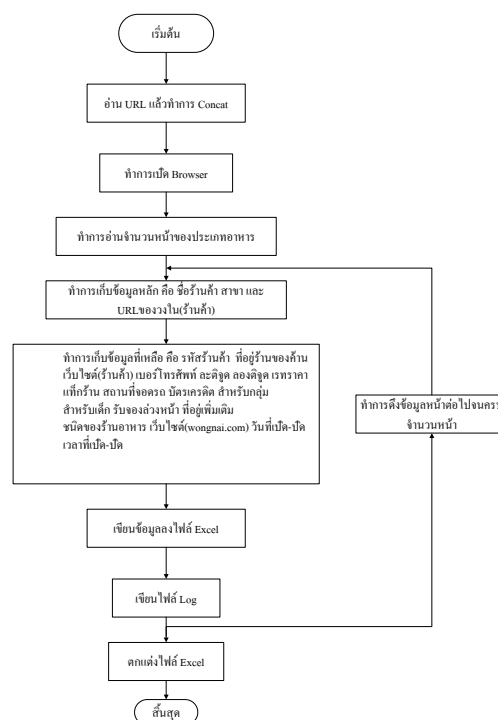
HtmlAgilityPack เป็น Library ที่มีไว้ช่วยในการทำงานกับหน้าเว็บไซต์ ซึ่งสามารถดึงข้อความจากเว็บไซต์ การได้ Node ลงมาเพื่อดึงข้อมูลจากส่วนนั้น การดึงข้อมูลโดยใช้ Xpath หรือ Innertext ได้

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

ชุดเครื่องมือที่ใช้สำหรับทดสอบเว็บแอปพลิเคชันอัตโนมัติ โดยประกอบด้วยเครื่องมือ 4 เครื่องมือ การใช้งานจะขึ้นอยู่กับวัตถุประสงค์ของการทดสอบในแต่ละองค์กร

3. วิธีดำเนินการทดลอง

การจัดทำโปรแกรมดึงข้อมูลจากเว็บไซต์ ในส่วนของผู้ใช้งาน
โปรแกรมนี้ได้ออกแบบส่วนต่อประสานผู้ใช้ให้ใช้งานง่าย และใน
ส่วนการทำงานของโปรแกรมจะทำงานโดยอัตโนมัติ โดยผู้ใช้งาน
โปรแกรมสามารถกรอกข้อมูลที่ต้องการจะดึงจากหน้าเว็บไซต์
หลังจากนั้นโปรแกรมจะทำการอ่าน URL และทำการเปิด
Browser หลังจากนั้นทางผู้ใช้งานได้ทำการเลือกประเภทอาหาร
แล้ว โปรแกรมจะทำการเปิดเข้าสู่หน้าประเภทอาหาร ต่อมาทำ
การอ่านจำนวนหน้าทั้งหมดของประเภทอาหาร แล้วทำการเก็บ
ข้อมูลหลัก ซึ่งหลังจากที่ทำการเก็บข้อมูลหลัก โปรแกรมจะทำ
การวนลูปเพื่อเข้าเก็บข้อมูลที่เหลือ ซึ่งแต่ละหน้าจะทำการวนลูป
จำนวน 20 รอบ เนื่องจากแต่ละหน้ามีร้านอาหารอยู่จำนวน 20
ร้าน และทำการเขียนข้อมูลลง Excel และเขียนไฟล์ Log
หลังจากนั้นทำการเก็บข้อมูลในหน้าต่อไปจนครบ และทำการ
ตกแต่งไฟล์ Excel แล้วทำการปิด Browser โดยมีขั้นตอนดัง
ภาพที่ 1

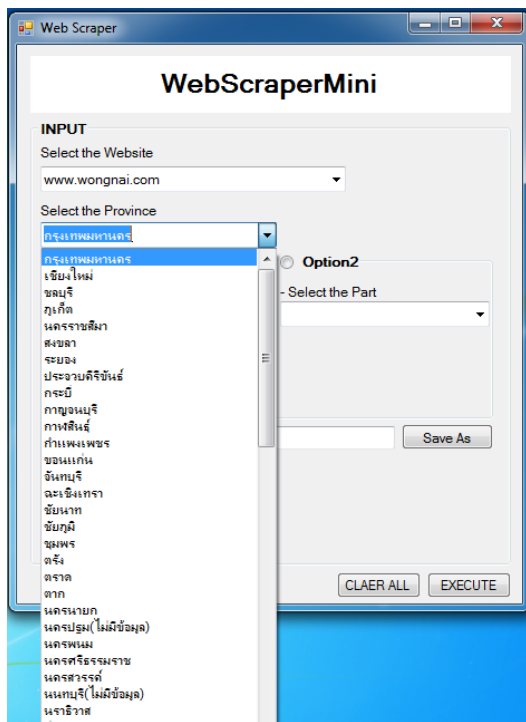


ภาพที่ 1 Flowchart ในส่วนของการ Process

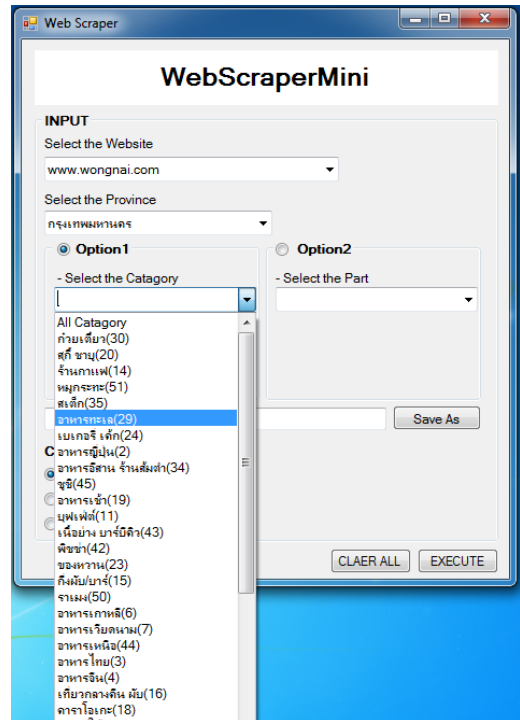
4. ผลการทดลอง

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

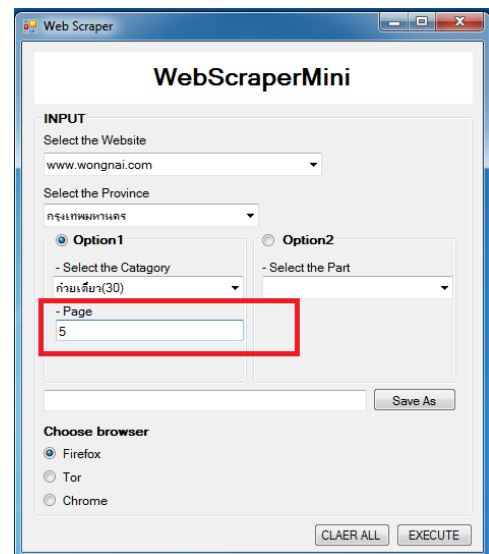
โปรแกรมดึงข้อมูลจากเว็บไซต์ออกเป็นไฟล์ Excel แบบอัตโนมัติ
จะทำการดึงข้อมูล รหัสร้านค้า ชื่อร้านค้า ที่อยู่ของร้านค้า
เว็บไซต์ เบอร์โทรศัพท์ ละติจูด ลองจิจูด เทรทราคา แท็กชื่อค้นหา
ร้าน สถานที่จอดรถ บัตรเครดิต สำหรับกลุ่ม สำหรับเด็ก รับจอง
ล่วงหน้า ที่อยู่เพิ่มเติม สาขา ประเภทอาหาร ชนิดร้านอาหาร
เว็บไซต์(wongnai.com) วันที่เปิด-ปิด เวลาที่เปิด-ปิด ดังภาพที่
2 - 8



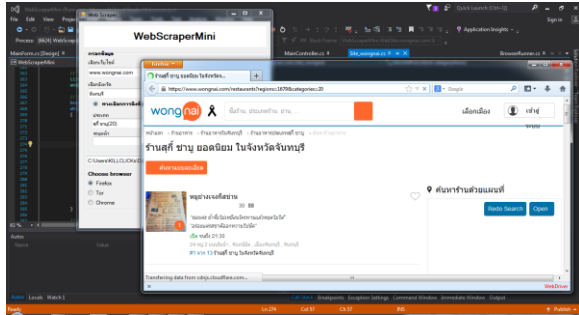
ภาพที่ 2 ทำการเลือกจังหวัด



ภาพที่ 3 เลือกประเภทที่ต้องการดึงข้อมูล



ภาพที่ 4 สามารถกรอกหน้าที่ทำการเริ่มบันทึกได้ หลังจาก
ข้อมูลครบแล้วทำการกด Execute เพื่อทำการดึงข้อมูล



ภาพที่ 5 ตัวอย่างการรันข้อมูลและทำการดึงข้อมูล

1	Id	Name	Street_Address	Website
2		ร้านข้าวต้มสุก	ช.ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
3	130254	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
4		ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
5	146279	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
6	166286	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
7	23460	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
8	13952	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
9	16914	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
10		ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
11	24090	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
12	185901	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
13	15209	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
14	181982	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
15	8324	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
16	7790	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
17	7813	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
18	178505	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
19	4414	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
20	118963	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
21		ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
22	10473	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
23	26168	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
24	158023	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
25	155122	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co
26	115080	ร้านข้าวต้มสุก	ถนน ศิริมงคลจากร 13 ถ.ศิริมงคลจากร	https://www.facebook.co

ภาพที่ 6 ตัวอย่างผลลัพธ์ไฟล์ Excel

PhoneNO	Latitude	Longitude	ResLocate
088-798-7996, 088-541-6646	18.7933772344101	98.97134599140168	ร้านข้าวต้มสุก
084-949-2828	18.79423839198281	98.97243154844477	ร้านข้าวต้มสุก
089-431-4040	18.7895917	98.973847	ร้านข้าวต้มสุก
088-261-1007	18.78960425594696	98.96259967761061	ร้านข้าวต้มสุก
093-935-4869	18.7976735808167	98.96226581186056	ร้านข้าวต้มสุก
08-1648-8238	18.79667788706477	98.96586542518332	ร้านข้าวต้มสุก
053-894-881	18.79801200068058	98.96927060061171	ร้านข้าวต้มสุก
053-221-921	18.794536	98.9696947	ร้านข้าวต้มสุก
085-030-3993	18.80155766673737	99.01074271576022	ร้านข้าวต้มสุก
053-213-284	18.7995773928679	98.972097754631	ร้านข้าวต้มสุก
094-628-4412, 081-980-6085	18.794678	98.953993	ร้านข้าวต้มสุก
053-211-765, 086-911-0022	18.8046672056609	98.96676209151745	ร้านข้าวต้มสุก
095-594-7443	18.807751	98.965894	ร้านข้าวต้มสุก
053-226-379	18.795231722546834	98.96678918956659	ร้านข้าวต้มสุก
081-472-9619	18.796835	98.968373	ร้านข้าวต้มสุก
053-809-129	18.789527	98.965666	ร้านข้าวต้มสุก
053-216-696	18.79385	98.971285	ร้านข้าวต้มสุก
081-764-6883	18.78864178567344	98.95955425396733	ร้านข้าวต้มสุก
081-5946597	18.795717294842127	98.985715889549	ร้านข้าวต้มสุก
084-500-1047, 089-701-8057	18.77633819393698	98.99910058626208	ร้านข้าวต้มสุก
053-248-502	18.787017	99.009737	ร้านข้าวต้มสุก
085-811-6335	18.793406803729486	98.93289915211492	ร้านข้าวต้มสุก
053-244405	18.781206418229303	99.00598762946515	ร้านข้าวต้มสุก
081-527-3343	18.792507	98.956259	ร้านข้าวต้มสุก
05-371-1663, 08-3208-7097	18.785338	98.983307	ร้านข้าวต้มสุก

ภาพที่ 7 ตัวอย่างผลลัพธ์ไฟล์ Excel

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	16:30:21	Catagory (30)	Page (17)	= success

ภาพที่ 9 ตัวอย่างไฟล์ Log

5. สรุปผลการทดลองและข้อเสนอแนะ

โครงการนี้จัดทำแอปพลิเคชันการดึงข้อมูลจากเว็บไซต์ออกเป็นไฟล์ Excel แบบอัตโนมัติ โดยใช้เทคนิคตามโครงสร้างของเว็บไซต์ โดยใช้เว็บไซต์ดวงใจ (www.wongnai.com) เป็นซึ่งเป็นเว็บไซต์บริการเครือข่ายสังคมของประเทศไทยที่มีความน่าเชื่อถือและข้อมูลที่มีการอัปเดตอยู่เสมอ เป็นกรณีศึกษา โดยแอปพลิเคชันจะทำการดึงข้อมูล เช่น ชื่อร้านค้า ที่อยู่ของร้านค้า เว็บไซต์ เบอร์โทรศัพท์ ละติจูด ลองจิจูด ประเภทอาหาร ชนิดร้านอาหาร เว็บไซต์ วันที่เปิด-ปิด เวลาที่เปิด-ปิด และข้อมูลอื่นๆ ที่เกี่ยวกับร้านอาหารมาเก็บลง Excel ซึ่งประโยชน์ที่ได้รับจากโครงการนี้คือจะช่วยลดทั้งค่าใช้จ่ายและแรงงานในการรวบรวมข้อมูลร้านอาหารที่อัปเดตอยู่เสมอ และประโยชน์ที่นำไปใช้ต่อสามารถนำข้อมูลนี้ไปอัปเดตให้กับหน่วยงานที่ใช้ข้อมูลจะทำให้ได้ข้อมูลปัจจุบัน

ข้อเสนอแนะคือถ้ามีการเปลี่ยนแปลงโครงสร้างของเว็บไซต์จะต้องมีการปรับปรุงการเขียนคำสั่งในโปรแกรม ซึ่งจะพัฒนาเทคนิคที่ใช้ได้กับทุกโครงสร้างต่อไป ซึ่งแนวทางในการพัฒนาต่อ นักพัฒนาโปรแกรมต้องสร้างเงื่อนไขเพื่อรองรับในการเปลี่ยนโครงสร้างของหน้าเว็บไซต์ ซึ่งโครงสร้างของเว็บไซต์สามารถเปลี่ยนแปลงได้ตลอดเวลา ซึ่งจะต้องสร้างเงื่อนไขออกมาเพื่อรองรับให้ใช้ได้ทุกกับโครงสร้าง

เอกสารอ้างอิง

- [1] กิตินันท์พลสวัสดิ์ .Visual C# 2010. พิมพ์ครั้งที่ 1 นนทบุรี : สำนักพิมพ์ไอทีซี, 2554.
- [2] เกรียงศักดิ์ จันทน์นอกการเขียนโปรแกรมทางธุรกิจ .1. พิมพ์ครั้งที่1 : มหาสารคามอภิชิตการพิมพ์, 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.