

Application of Web Scraping and Google API Service to Optimize Convenience Stores' Distribution

Quang Thai LE*, Davar PISHVA**

*Faculty of International Management, Ritsumeikan Asia Pacific University, Beppu, Japan

** Faculty of Asia Pacific Studies, Ritsumeikan Asia Pacific University, Beppu, Japan

quanle11@apu.ac.jp, dpishva@apu.ac.jp

Abstract— In the current competitive market, offering a relatively cheaper price for the commodity plays an important role in obtaining a greater share of the market for a corporate entity as it encourages more customers to purchase from its product. Convenience stores are becoming indispensable for the Japanese society, there are presently 54,008 retail stores throughout Japan and their number continues to increase. The reasons behind such popularity are convenient locations, attractive products and long trading hours. Their price, however, are more expensive than regular supermarkets and groceries stores mainly due to their numerous overhead cost, the major of which is their requirement for refurbishment of goods several times per day. In this study, a practical approach that utilizes computer techniques to find an optimal vehicle routing scheme for goods and service delivery to multiple convenience stores is investigated. Although many scholars have already investigated the location problem of supply chain facilities and centers under different conditions, this study takes a programming approach using Web Scraping and Excel VBA and hope to turn it into a cheap but powerful Excel Add-in module or real-time navigation function. A mathematical 'network flow model' is initially developed to examine the problem. Geographical data of convenience stores, their associated warehouses, garbage dumpsites and gas stations are subsequently retrieved through programming with the 'web scraping' technique. A computer program that utilizes Google API service is then developed to solve the optimal networking problem. Validity of obtained results is also examined by other known method to justify its optimality and fast performance.

Keywords— Supply Chain, Maximal Covering, Assignment, Modeling, Optimization, Web Scraping, Google API Service

I. INTRODUCTION

Convenience stores play an important role in the Japanese society. There are totally 54,008 convenience stores located throughout Japan, generating a total sale of 9.81 trillion yen for the fiscal 2013 [1]. Convenience stores are attracting Japanese customers because of their convenient locations and supply of ready-to-eat and fresh foods [2], contributing 36% of the Japan Retail Food Market in 2012 [3-4]. However, in order to meet huge demand of highly dense Japanese population, convenience store companies have to deliver goods to their retail stores several times per day, which

ironically causes a big overhead cost. Such overhead cost is usually associated with the transportation time, fuel refill expenses, and the amount of energy fuel consumed for the delivery, not to mention the associated adverse environmental impacts. This paper addresses the above problem under an optimization perspective, and demonstrates a practical programming approach for the problem.

II. CONVENIENCE STORE DISTRIBUTION PROBLEM

On a daily basis, convenience retailers have to deliver goods by trucks from a distribution center to several stores, collect all the garbage, fill-up the gasoline if necessary, and dump the collected garbage at a designated dumpsite. On the average, replenishments of goods for a single convenience store are carried out 9 times a day [5] mainly because of the small size of the store. As there exists many routes for such deliveries and garbage dumping, a systematic approach to handle the needs of a network of convenience stores would help in achieving less fuel cost, less transporting time, and also reduce the amount of carbon-dioxide emission to the environment.

In this research, a computer program is created and introduced to scrape geographical data of over 97,000 convenience store and gas station locations in Japan, and find an optimal transportation route to deliver goods from a warehouse (distribution center) to up-to 8 convenience stores in a given vicinity, pass by a gas station for vehicle refuelling, and end up at a garbage dump site. Moreover, a Microsoft Excel Add-in module is created to carry out the implemented algorithm directly in the Excel. This practical approach to the convenience stores' distribution optimization problem can help save fuel cost, shorten the delivery time, reduce adverse environmental impacts and ultimately lower product cost.

III. METHODOLOGY

The general approach employed in this research is quantitative and the following four schemes are used in formulating, analysing and validating its obtained results.

A. Network Flow Modeling

The convenience stores' distribution optimization problem is examined as a "Network Flow Problem" by assuming each

of physical locations in the distributing routes as a node in the network, distance and travelling duration as weights of arcs connecting the nodes. The network's optimization objective is therefore to find a particular flow sequence through the nodes that yields in optimal transportation time [6-15], a mathematical representation of which is shown in Fig. 1.

Notation for each node:

$$a_1, a_2, a_3, \dots, a_n$$

Arc weight between node:

$$a_i \text{ and } a_j: \text{Weight}(a_i, a_j)$$

Whether exist a flow from node a_i to node a_j :

$$\text{Flow}(a_i, a_j) = 1(\text{exist}) | \text{Flow}(a_i, a_j) = 0(\text{non-exist})$$

Minimize:

$$z = \sum_{i=1}^n \sum_{j=1}^n \text{Flow}(a_i, a_j) \text{Weight}(a_i, a_j)$$

Fig. 1. Mathematical representation of Network Flow Model

In order to solve this Network Flow optimization problem, collecting information about geographical locations of convenience stores, gas stations, warehouses and garbage dumpsites, as well as the distance and travel time between these locations are indispensable.

B. Data Collection Using Web Scraping

Web scraping is a programming technique that extracts data from the World Wide Web. This technique equips the program with an artificial intelligence, helps it to continuously surf the Internet and extract relevant pieces of electronic data.

Web scraping can be practiced by a variety of programming languages that support Hypertext Transfer Protocol (HTTP) programming. Ruby programming language and its Nokogiri gem (Ruby external library) is used in this research to retrieve the convenience business' relevant data because of its high performance and HTTP parser libraries availability. As Nokogiri library supports several document encodings (including UTF-8) and parser methods (XPath or CSS3 selector) [16], the Ruby gem can help in effectively retrieving data from a variety of online sources.

This research mainly relies on NAVITIME website (<http://www.navitime.co.jp>) which contains a list of convenience stores and gas stations throughout Japan in "HTML" format with UTF-8 encoding, together with their complete address, telephone number, and service details. Since the user interface of this website is designed for interactive usage, several clicks are needed in order to retrieve information about a single physical location and each click usually opens a new page. Hence a programming technique is employed to simultaneously retrieve all the required data a pseudo algorithm of which is shown in Fig. 2.

In order to speed up experimental analysis, one can save all the extracted data in a local CSV (Comma Separate Values) file and use it as an input data during the optimization analysis.

Result: CSV Database File

Access the list of all convenience stores web page;

repeat

repeat

Access each store's link;

Extract the store information;

Append the extracted data into the result CSV file;

Return to the previous list page;

until The last store of the list;

Access the next list page;

until The last page list;

Fig. 2. Web scraping algorithm

Table 1 shows the time it takes to extract all of the 56,925 convenience stores and 39,660 gas stations throughout Japan using the above technique.

TABLE 1. EXTRACTION TIME OF CONVENIENCE STORES AND GAS STATIONS USING RUBY SCRIPT AND NOKOGIRI LIBRARY

Data	Convenience Store	Gas Station
No. of Locations	56,925	39,660
Extraction Time	96 minutes	80 minutes
CSV File Size	16.3 MB	14 MB

C. Google API Service and Computer Algorithm

1) Google API Service: Google API (Application Programming Interface) is a well-known web service that provides online tools for developers who utilize Google's data and infrastructure. Google Direction API is one such service that helps retrieve geographical distance and travelling time from one physical location to another. Moreover, the service also helps to calculate the optimal route in terms of time to travel through a network of several physical waypoints (locations).

Google Direction API provides freemium service that lets the user to make up to 8 complementary waypoints per request, and 2,500 requests per 24-hour period free-of-charge. Additionally, the Google API for work customers (payment required) allow the user to make a maximum of 100,000 directions requests per day, and up-to 23 waypoints in each request [17]. Such functionality of the Google Direction API is quite useful for both testing and practical use in solving the convenience store's distribution problem.

In the implementation process, for each distribution network, a HTTP request is sent to a designated address (<http://maps.googleapis.com/maps/api/directions>) with some parameters including:

- **Origin point:** The location from where the trip starts.
- **Via points:** A list of all geographical points to pass-by during the trip.

- **Destination point:** The final destination where the trip ends.

The response from Google API is returned in either JSON or XML format that can be processed to get the desired optimal traveling order. If delivery requires refuelling along the way, several additional processes related to the associated information of the near-by gasoline stations are carried out. At the end of the process, among all the possible alternative routes, the algorithm returns the one that provides the shortest possible travel time.

2) Computer Algorithm: A computer program is written in Ruby language to accomplish the process of retrieving pertinent data and transmitting them to Google API service for an optimal routing calculation service. The program is built to interact with users in a friendly manner so as to obtain user's instruction prior to solving the problem.

In order to come up with an optimal route for the convenience stores' distribution problem, the program automatically carries out in the following operations:

Request for Input Parameters

A user of the system needs to supply the following program input parameters:

- Location of **warehouse (distribution center)**,
- Location of **garbage dumpsite**,
- Delivery area (**postal code**),
- Intention on vehicle **refuelling**.

Data Mining and Optimization Process

- Locate convenience stores and gas stations in the vicinity of the desired delivery area.
- Request the user to select up-to 8 (7 in case of vehicle refuelling) convenience stores from the extracted list.
- Forward an appropriate parameter list to the Google API service.

Display Output Result

Interpret responses that are retrieved from Google API service and display them in an easy to follow manner.

3) Optimality Validation: Because the program heavily depends on the Google Service API in finding the optimal route, an independent optimality validation test is carried out to justify the validity of obtained result.

For doing so, we have equipped the program with export data functionality, which enables a user to export all the retrieved data, including locations of the warehouse, convenience stores, gas stations and garbage dumpsite, as well as the distance and travel time through these locations, into Excel. An independent optimization analysis is then performed using the Solver Add-in function of Excel to verify the validity of the originally obtained solution.

D. Integration as Microsoft Excel Add-in

1) Microsoft Excel Add-in: Through Microsoft Excel Add-in, functionality of Excel can be extended by means of additional analytical functions most of which developed by 3rd parties. Solver is an example of such product, which is developed by Frontline Systems Inc. and can be used for numerous optimization and linear programming problems [18]. This research has implemented optimization of convenience stores' distribution as an Excel Add-ins and its prototype is written in VBA (Visual Basic for Applications), which can make direct use of the numerous retrieved data.

The Add-in enables a user to look up for all convenience stores located in a given geographic area (specified by a postal code) and find the optimal route for the convenience stores' distribution. As shown in Figs 3-4, through its GUI (Graphical User Interface) users can select specific cells as the input and output data location directly from the Microsoft Excel working space. Furthermore, the Add-in also automatically imports the CSV database file including the retrieved information of convenience stores and gas stations into new worksheets namely, "Convenience Stores" and "Gas Stations", for further usage.

2) Convenience Stores Look Up Function: The Add-in makes use of the imported "Convenience Stores" and "Gas Stations" worksheets and VBA's "Find" function to list all of the convenience stores that are located in the vicinity of an area which is specified by a user through an area postal code. Fig. 3 shows Add-in's GUI for the input parameters of the look up function.

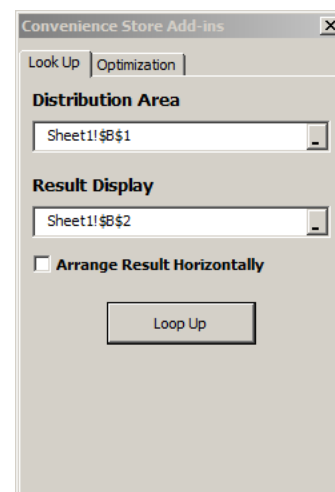


Fig. 3. Add-in's GUI for the input parameters of the look up function

3) Distribution Route Optimization Function: This function retrieves the optimal distribution route for a specific network of up-to 8 convenience stores (7 in case of vehicle refuelling) and Fig. 4 shows the Add-in's GUI for its required input parameters.

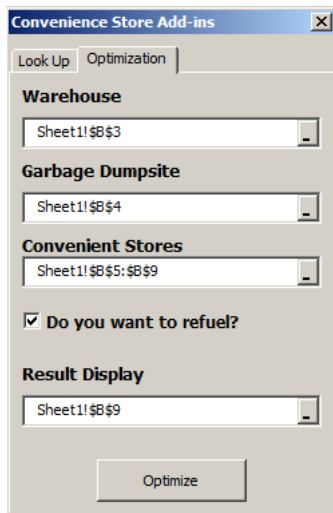


Fig. 4. Add-in's GUI for the input parameters of the optimization function

The function uses the external Microsoft XML (version 6.0) library in VBA to send HTTP request to Google Direction API service with the geographic data (convenience stores' addresses) collected from the Microsoft Excel worksheet, and then extract the optimal travelling sequence from the response in XML format.

IV. RESULTS AND DISCUSSION

This section briefly explains the obtained results and discusses validity and limitation of the approach.

A. Algorithm Testing

In order to test correct operation and reliability of a newly developed program, we initially used it to gather various data on roughly 56,925 convenience stores and 39,660 gas stations that are located throughout Japan. We then focused our analytical modeling approach to convenience stores that are located in Oita prefecture. Our finding shows that out of 1,765 different districts in Oita prefecture that have a unique postal code, (a complete list of which can be accessed at the link <http://homepage1.nifty.com/tabotabo/pzips/oita.htm>), in 754 or 42% of them there exist convenience stores. We then created examples of convenient store networks with up-to 8 stores per network for supply and garbage collection services. In about 95% of them, the program could come up with an optimal routing solution in less than 1 minute.

B. Accuracy and Performance

In general, routing optimization problem requires much time and computational resources in order to come up with a feasible solution. In this particular case, since optimization problem consisted of a network having a warehouse, up to 8 convenience stores, and a garbage dumpsite, our algorithm had to examine a total of 40,320 possible routes in determining an optimal solution that could yield the least travelling duration or distance. Even though there are many proven algorithms, such as Evolutionary algorithm, to solve the routing optimization problem [19], their analysis in general takes much more time.

This was realized when we validated optimality of our result with that of Evolutionary algorithm built into Excel's Solver Add-in function [18] using its default settings shown in Table 2.

TABLE 2. DEFAULT PARAMETERS OF EXCEL'S SOLVER ADD-IN' FUNCTION

Convergence	0.0001
Mutation Rate	0.075
Population Size	100
Random Seed	0
Maximum Timeout without Improvement	30

Although both methods yielded the same result in terms of optimality, performance of our approach which relies on Google API is much faster than that of the Evolutionary algorithm built into Excel's Solver Add-in function, as shown in Table 3. Despite the fact that free version of Google API service is limited to a network of 10 nodes, its enterprise version can handle a network of up to 25 nodes.

TABLE 3. PERFORMANCE COMPARISON BETWEEN GOOGLE API SERVICE AND EVOLUTIONARY ALGORITHM OF EXCEL SOLVER ADD-IN

Testing Network	Evolutionary Algorithm by Excel Solver Add-in (approximately)	Google API service
Network of 6 locations	42 seconds	Less than 10 seconds
Network of 7 locations	51 seconds	Less than 10 seconds
Network of 8 locations	52 seconds	Less than 10 seconds
Network of 9 locations	53 seconds	Less than 10 seconds
Network of 10 locations	58 seconds	Less than 10 seconds

C. Excel Add-in Implementation

The originally developed Add-in can be easily installed and used to find the optimal distribution route with any version Excel. This means that the convenience retailing owners can utilize the developed algorithm together with an already well-established database stored in Excel worksheets.

Fig. 5 shows a sample Excel worksheet design that can utilize the Add-in to calculate the optimal route for a convenience stores' distribution network.

	A	B
1	Warehouse	別府駅
2	Garbage Dumpsite	別府駅
3	Postal Code	8740919
4		
5	No	Look Up Result
6	1	大分県別府市北浜3丁目818-166
7	2	大分県杵築市大字杵築字北浜700番1-2
8	3	大分県杵築市大字杵築字北浜665-29
9	4	大分県別府市石垣東7丁目2055
10	5	大分県別府市石垣東3丁目5-15
11		
12	Convenience Store Selection	
13	1	大分県別府市北浜3丁目818-166
14	3	大分県杵築市大字杵築字北浜665-29
15	4	大分県別府市石垣東7丁目2055
16	5	大分県別府市石垣東3丁目5-15
17		
18	Optimal Distribution Order	
19	Warehouse	別府駅
20	1	大分県別府市石垣東3丁目5-15
21	2	大分県別府市石垣東7丁目2055
22	3	大分県杵築市大字杵築字北浜665-29
23	4	大分県別府市北浜3丁目818-166
24	Garbage Dumpsite	別府駅

Fig. 5. A sample Excel worksheet design for calculating convenience stores' distribution route. Green values are inputted by the user, red ones are outputs generated by the Add-in, the blue ones are results from Excel's Vlookup function.

V. FUTURE RESEARCH

A. Online Network for Real-time Navigation, Data Collection and Pattern Learning

Although the recommended approach provides a practical solution for the convenience stores' distribution problem with the goal of optimizing the delivering time, which would result in reduction of gas consumption; the problem can be also examined under a pattern learning perspective to have an inner sight into the convenience stores' distribution process.

Specifically, apart from travelling time, there are also many other factors that contributes to the gas consumption during the product delivering process, namely; weather, temperature, route terrain, vehicle driving speed, etc. However, not all such factors can be easily controlled and calculated. Hence, a promising approach is to collect data on such factors' impacts on the volume of gas consumption and apply machine learning techniques so as to understand the patterns of how much each of these factors could contribute to the actual gas consumption. Then, an algorithm that also takes these factors into consideration can be developed. Moreover, because the data is expectedly retrieved in real-time, the learning pattern is continuously updated to ensure the most pragmatic suggestion.

To obtain this, the research promotes further examination into not only pattern learning techniques, but also ubiquitous & sensor network for real-time data collection.

B. Microsoft Excel Add-in Extension

Considering the fact that Microsoft Excel is presently the most popular digital planning tool, the introduced Add-in has the potential to be extended to support a wide range of specific distribution problems or scenarios, not only in convenience store business but also in other distribution related fields. Specifically, the Add-in can be extended as a real-time data synchronization tool to retrieved data directly into worksheet from a server or sensor network that measures and collects the convenience stores' delivery process related variables such

as average travel speed, actual travel time duration, vehicle's gas consumption, etc. Capturing such data into the Excel worksheet will allow several practical and complementary data analyses.

VI. CONCLUSION

The research promotes the practice of utilizing available technology and services to solve today's pragmatic supply chain management problem. Particularly, a practical approach to convenience stores' retailing distribution optimization problem is introduced. The approach not only recommends the utilization of Google API service to increase performance of the solution to the problem, but also the usage of "web scraping" programming technique to easily collect huge amount of relevant data available on the Internet. Considering its reliability and fast performance, the approach can be extended for practical applications. The approach can also be extended to solve similar supply chain management in other business fields that require regular delivery services.

REFERENCES

- [1] Nikkei Asian Review. (2014, July 23). Japan's top convenience stores dominate like never before. Retrieved from <http://asia.nikkei.com/Business/Consumers/Japan-s-top-convenience-stores-dominate-like-never-before>
- [2] NIPPONIA No.19. (2011, December 15). Convenience Stores and the Japanese Shopper. Retrieved from <http://web-japan.org/nipponia/nipponia19/en/feature/feature05.html>
- [3] USDA Foreign Agriculture Service. (2013, December 12). Japan Retail Food - Retail Sector 2012.
- [4] METI. (2014, October 14). Monthly Report on the Current Survey of Commerce. Retrieved from http://www.meti.go.jp/english/statistics/tyo/syoudou_kakuho/index.html
- [5] 7-Eleven. (n.d.). Convenience Store Operations. Retrieved from http://www.7andi.com/dbps_data/_template/_user/_SITE/_localhost/_res/en/ir/library/co/pdf/p24_33.pdf
- [6] Peltier Technical Services, Inc. (n.d.). Using Solver in Excel VBA. Retrieved from <http://peltiertech.com/Excel/SolverVBA.html>
- [7] Managerial Decision Modeling (International Edition) Cliff T. Ragsdale (Thomson South-Western 2011/01/01)
- [8] Data Analysis and Decision Making with Microsoft Excel Wayne L. Winston, S. Albright, Christopher J. Zappe (South-Western 2010/12/04)
- [9] VBA for Modelers: Developing Decision Support Systems With Microsoft Office Excel S. Christian Albright (South-Western Pub 2011/08/26)
- [10] Management Science: The Art of Modeling with Spreadsheets Stephen G. Powell, Kenneth R. Baker (Wiley 2010/10/26)
- [11] Structured Decision Making: A Practical Guide to Environmental Management Choices Robin Gregory, Lee Failing, Michael Harstone, Graham Long, Tim McDaniels, Dan Ohlson (Wiley-Blackwell 2012/03/19)
- [12] Decision Making in Natural Resource Management: A Structured, Adaptive Approach Michael J. Conroy, James T. Peterson (Wiley-Blackwell 2013/04/01)
- [13] Optimization Modeling with Spreadsheets Kenneth R. Baker (Wiley 2011/05/24)
- [14] Handbook of Decision Analysis (Wiley Handbooks in Operations Research and Management Science) Gregory S. Parnell, Terry Bresnick, Steven N. Tani, Eric R. Johnson (Wiley 2013/01/24)
- [15] Data Driven Business Decisions (Statistics in Practice) Chris J. Lloyd (Wiley 2011/10/25)
- [16] Tutorials - Nokogiri. (n.d.). Retrieved from <http://nokogiri.org>
- [17] Google Developers. (n.d.). The Google Directions API - Google Maps API Web Services. Retrieved from <https://developers.google.com/maps/documentation/directions/>
- [18] Frontline Systems Inc., Solver URL: <http://www.solver.com>
- [19] Kylie Bryant, 1., & Arthur Benjamin. (n.d.). Genetic Algorithms and the Traveling Salesman Problem.