ICIC Express Letters
Part B: Applications
Volume **15**, Number **3**, March **2024**

ICIC International ©2024 ISSN 2185-2766
pp. 287–293

VOC ANALYTICS BASED ON WEB CRAWLING AND ROBOTIC PROCESS AUTOMATION

Chang Seong Ko, HaeKyung Lee and Taioun Kim*

Department of Industrial and Management Engineering
Kyungsung University
309, Suyeong-ro, Nam-gu, Busan 48434, Korea
{ csko; hshklee72 }@ks.ac.kr; *Corresponding author: twkim@ks.ac.kr

Received June 2023; accepted September 2023

Abstract. *Recently, by the help of the emerging technologies such as AI, big data, cloud computing, and IoT, the previous human labor is being replaced by digital labor. Robotic Process Automation (RPA) is a software technology that transforms simple, repetitive tasks performed by humans into automated process. In recent years, offline commerce is rapidly changing to online due to the corona pandemic and changes in the lifestyle. An important factor in decision-making in online commerce is the Voice of Customer (VoC). However, VoCs are very difficult to process manually due to the volume, diversity and frequency of data. Therefore, it is a very effective method to perform data collection, analysis, processing, and feedback on VoCs using RPA. This study aims to optimize customer service by automatically collecting customer opinions related to online shopping using Web crawling technology. The gathered big data is analyzed and processed using sentence tokenizing, topic modeling, sentiment analysis, and Word Cloud technology. The result of VoC analytics will help to improve the customer satisfaction.*

Keywords: Robotic Process Automation (RPA), Web crawling, Voice of Customer (VoC), Information extraction, UiPath

1. **Introduction.** Robotic Process Automation (RPA) is a software that automatically follows the simple actions of a person working with a computer as a semi-bot, and is an automation technology widely used for automation of factory diversification and routine office work. The advantages of RPA are as the following. Human resources can be made more efficient by replacing simple, repetitive and low value-added tasks. Quality can be improved by reducing errors that often occur in manual work. Productivity can be increased by investing time spent on simple tasks into higher value-added tasks.

For the social background, the need to improve productivity in office work has increased due to the reduction of working hours, the decrease in the workforce due to the low birth rate and aging population, and the need to strengthen corporate competitiveness. In addition, as AI technology enables low system introduction cost and high ROI for IT infrastructure, it is becoming possible to support decision-making in professional fields. RPA is a representative technology for performing such a digital labor.

In general, the evolutionary stages of automation can be divided into multiple stages: macro process, RPA process, process integration, intelligent automation, and autonomous intelligence; therefore, RPA is still in its infancy. As AI is being developed rapidly in recent years, all automation processes will develop in convergence with AI.

RPA is an umbrella term for tools that operate on the user interface of other computer systems in the way a human would do. RPA aims to replace people by automation done in an "outside-in" manner. This differs from the classical "inside-out" approach

DOI: 10.24507/icicelb.15.03.287

to improve information systems. Unlike traditional workflow technology, the information system remains unchanged [1].

For example, many things done in the office include simple repetitive work of simply printing out customer request data and sending it by e-mail, and creative work of planning, writing reports, and making decisions based on this. Also, there are complex, complicated and challenging works which require creative and break-through approach. The former is the realm of RPA, while the latter is the realm of Cognitive Intelligence (CI). The two types of tasks and areas are shown in Figure 1.

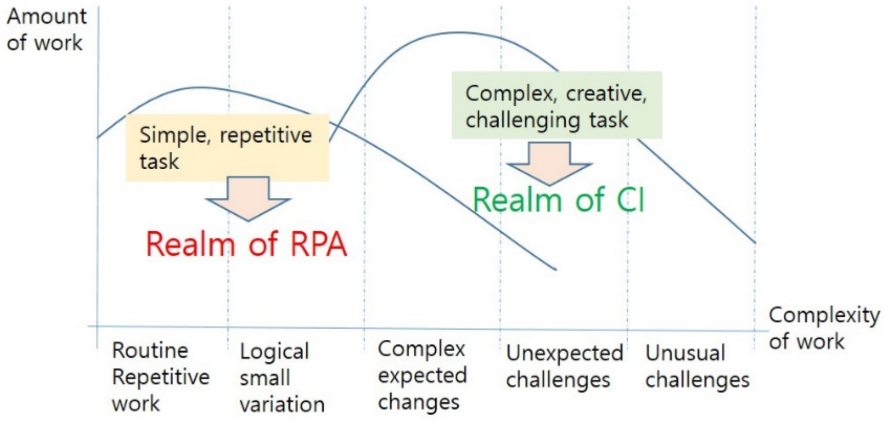


Figure 1. Simple/repetitive work and complex/creative tasks

Web crawling is a process of automating the extraction of data in an efficient and fast way. With the help of Web crawling, you can extract data from any website, no matter how large is the data, on your computer.

UiPath automation tool is used for extracting Web page data and produces data automatically in the CSV document [2]. This method is valuable in small scale industries.

In the affiliated colleges in India, the faculty need to enter the hall ticket number of each student every time to extract the data and then manually copy the data such as hall ticket no, name and grades of each subject along with the score from the websites mentioned above. To solve this monotone tasks, the examination results from the websites can be extracted automatically using the RPA technique [3].

One field where Web scraping is widely used is travel review analysis. Web scraping technology is utilized to analyze the data of TripAdvisor's social page, and uses it to visualize and analyze a lot of statistical data [4].

For smart online shopping, Web crawling and scraping methods are applied for identifying best deals from five e-commerce websites. For the methods, Python Requests and Selenium are adopted [5].

Understanding the customer's voice in a product is essential from both producer and consumer perspectives. Producers can understand customers' thoughts about their products, and consumers can know how to purchase products they want at the lowest price. In this respect, RPA and Web crawling technology are the optimal methodologies.

The purpose of this study is to propose a Web crawling and RPA model which analyze VoCs in the online website. Section 2 presents the methodology and model. In particular, we present a model that acquires customer voices through Web crawling and analyzes them with RPA technology at electronic product purchase sites where online transactions are frequent. In Section 3, the implementation according to the model is carried out, and in Section 4, conclusions and future research areas are presented.

2. **Methods.** Today, the VoC is very important due to the rapid increase in e-commerce and online shopping and the expansion of SNS influence. These VoC data are vast in

volume, occur frequently, and take the form of big data due to the diversity of data. In addition, the new shopping patterns of the MZ generation generate new trends in the shopping methods. Therefore, automating the collection and processing of VoC data is crucial to most online shoppers.

2.1. **VoC collection using Web crawling and analysis model.** Figure 2 shows the VoC collection and analysis model. This model consists of VoC collection and analysis processing model. First, through crawling of the web server, review data about interested product is collected and stored. In the next analysis step, first go through sentence tokenizing and then perform sentiment analysis by topic modeling. In addition, after executing sentence tokenizing, Pos Tagging is performed through natural language processing, and Word Cloud is implemented from this.

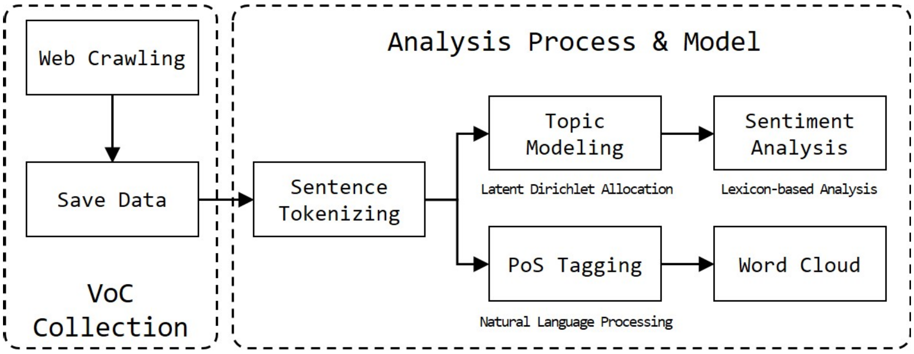


Figure 2. VoC collection and analysis model

When Web crawling, focused crawling is required due to the vast amount of data. [6] proposed classifiers for assigning any document retrieved from the Web to one of the layers of the merged context graph, and for quantifying the belief in such a classification assignment. A context graph represents how a target document can be accessed from the web. They use keyword indexing of each document using a modification of TF-IDF (Term Frequency-Inverse Document Frequency).

For a vocabulary V, The TF-IDF score $v(w)$ of a phrase w is computed using the following function:

$$v(w) = \frac{f^d(w)}{f^d_{\max}} \log \frac{N}{f(w)} \tag{1}$$

where $f^d(w)$ is the number of occurrences of w in a document d, $f^d_{\max}$ is the maximum number of occurrences of a phrase in a document d, N is the number of documents in the reference corpus and $f(w)$ is the number of documents in the corpus where the phrase w occurs at least once.

For quantifying and evaluating classifiers, the above index can be useful. However, the application domain is slightly different with our research, and we follow the framework shown in Figure 2.

2.2. **RPA model.** The term ‘Robotic Process Automation’ may connote views of physical robots wandering around offices helping human tasks, but it is a software solution. RPA has three distinctive features compared to other automation tools [7].

- RPA is easy to configure, so developers do not need programming skills.
- RPA software is non-invasive. There is no need to create, replace or further develop expensive platforms.
- RPA is enterprise-safe. It is a robust platform designed to meet enterprise IT requirements for security, scalability, auditability, and change management.

RPA is a technology that uses software robots to automate simple tasks that need to be repetitively handled by humans. The expected effects of RPA are as follows:

- Human resource efficiency;
- Quality improvement;
- Improved productivity by focusing on high value-added work.

Many applications of RPA include shared services. The shared service integrates simple and repetitive support functions such as IT, accounting, purchasing, and SCM performed in various departments so that a specific department can perform them effectively and efficiently. Shared services may provide lower cost, tighter control, higher service level, and scalability.

The RPA model adopted in this study consists of data acquisition through web crawling, real data analytics, and a dashboard that shows them. Data analytics consists of Topic Modeling, PoS Tagging, Sentiment Analysis, and Word Cloud. Dashboard also automatically creates and displays these contents.

2.3. **VoC analysis framework using RPA.** In this study, basic data is collected through Web crawling on the contents of online shopping customers' opinions on products. For the data collected in this way, in-depth analysis is conducted using text mining techniques. As mentioned earlier, technologies such as Latent Dirichlet Allocation, Lexicon-based Analysis, KoNLP and Word Cloud are utilized.

Based on the text mining process, RPA technology is applied. Expected results include various related statistical graphs and a Word Cloud representing customer preference. The framework of Web crawling and text mining for VoC analysis is shown in Figure 3.

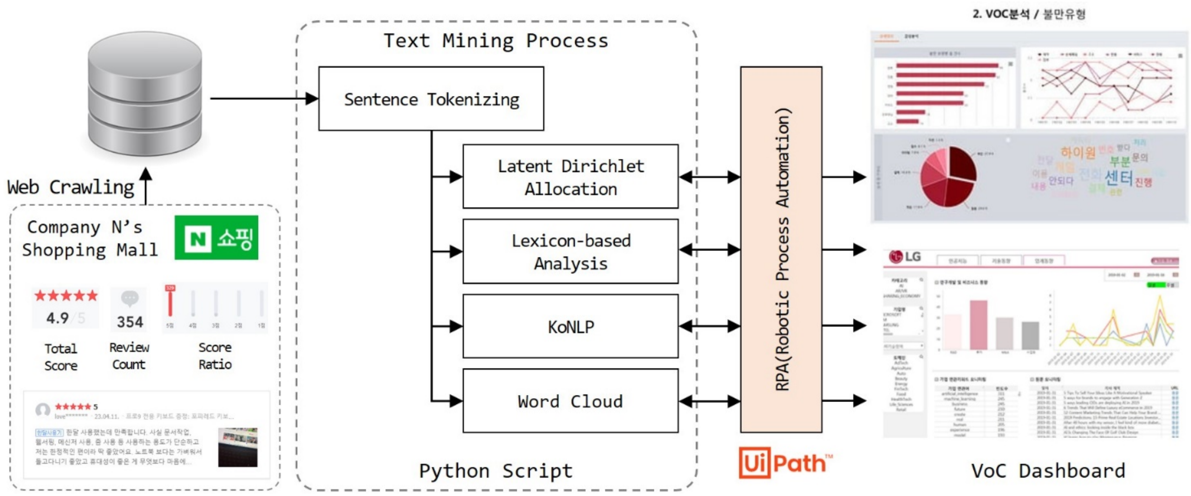


FIGURE 3. Framework of Web crawling and text mining for VoC analysis

A sentiment based rating prediction and recommendation model which is for predicting the rating of products from user reviews is presented in [8]. Emotion detection aims at detecting emotions like, happiness, frustration, anger, and sadness. in the reviews. Just like mining the opinion from the review emotions also has its importance to form precise sentiment about a product.

3. Implementation and Results.

Following the above framework, customer voice has been analyzed as follows.

3.1. **Web crawling.** For the implementation, most popular webpage is selected for analysis of customer review about a product. The product category is selected from the top portal shopping site in Korea: [naver] – [shopping] – [computer] – [notebook]. The product

Product code	Manufacturer	Model	Category1	Category2	Title	Registered date	Minimum price	Review frequency	Points

(a) Product data items

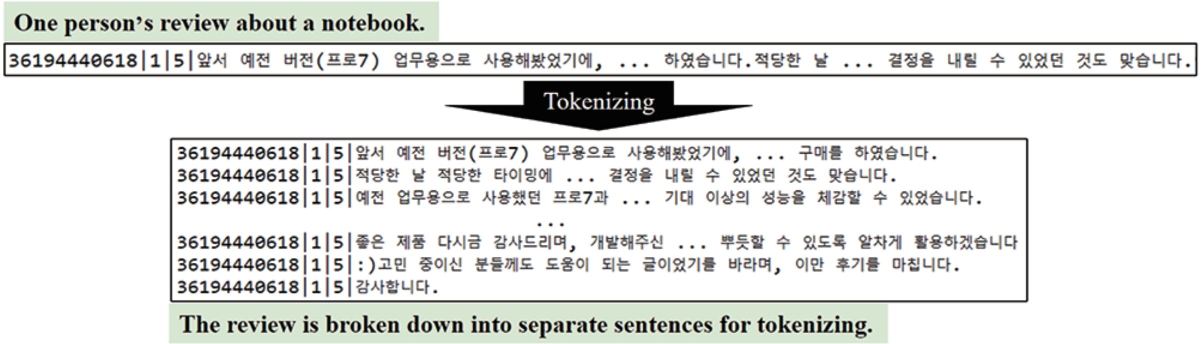
Product code	Review	Sequence	Points

(b) Review data items

Figure 4. Data structure for product and review used in Web crawling

data and review data field are composed of the following property. Figure 4 shows data structures for product data and review data.

3.2. **Sentence tokenizing.** Sentence tokenization is the process of splitting paragraphs and sentences into smaller units (i.e., sentence). The first step is to split a person's review into separate sentence. Then, tokenized sentence is used as an input for Word Cloud and sentiment analysis. Figure 5 shows the case of one review broken into a separate sentence.



One person's review about a notebook.

```
36194440618|1|5|앞서 예전 버전(프로7) 업무용으로 사용해봤었기에, ... 하였습니다.적당한 날 ... 결정을 내릴 수 있었던 것도 맞습니다.
```

Tokenizing

```
36194440618|1|5|앞서 예전 버전(프로7) 업무용으로 사용해봤었기에, ... 구매를 하였습니다.
36194440618|1|5|적당한 날 적당한 타이밍에 ... 결정을 내릴 수 있었던 것도 맞습니다.
36194440618|1|5|예전 업무용으로 사용했던 프로7과 ... 기대 이상의 성능을 체감할 수 있었습니다.
                                    ...
36194440618|1|5|좋은 제품 다시금 감사드리며, 개발해주신 ... 뿌듯할 수 있도록 알차게 활용하겠습니다
36194440618|1|5|:)고민 중이신 분들께도 도움이 되는 글이었기를 바라며, 이만 후기를 마칩니다.
36194440618|1|5|감사합니다.
```

The review is broken down into separate sentences for tokenizing.

Figure 5. Sentence tokenizing of a person's review

3.3. **Word Cloud.** Word Cloud or text cloud is a visualization of word frequency in a given text as a weighted list. The technique has recently been popularly used to visualize the topical content of interested domain. The result from the proposed model is shown as follows.

First, Table 1 shows the frequency of word that will appear in the Word Cloud.

Table 1. Summary of recognized data frequency for the Word Cloud

Ranking	Noun		Frequency	Ranking	Noun		Frequency
1	배송	Delivery	3,982	6	윈도우	Window	2,877
2	가격	Price	3,692	7	인치	Inch	2,745
3	성능	Performance	3,226	8	무게	Weight	2,379
4	설치	Installation	3,197	9	디자인	Design	2,100
5	화면	Screen	3,018	10	가성	False	1,655

Next, Word Cloud from the VoC can be extracted and displayed as Figure 6. It shows the popularity of words or phrases by making the most frequently used words appear larger or bolder compared with the other words around them. A Word Cloud contains a set of data such as a list of words, text from a blog post, or a collection of written items like a series of articles.

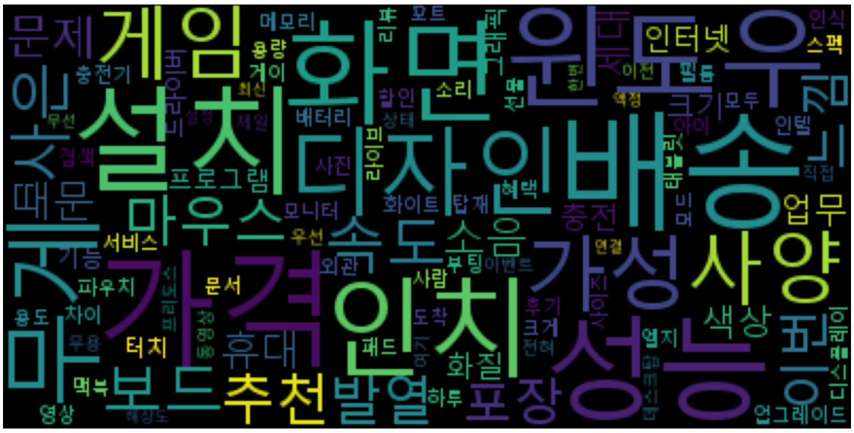


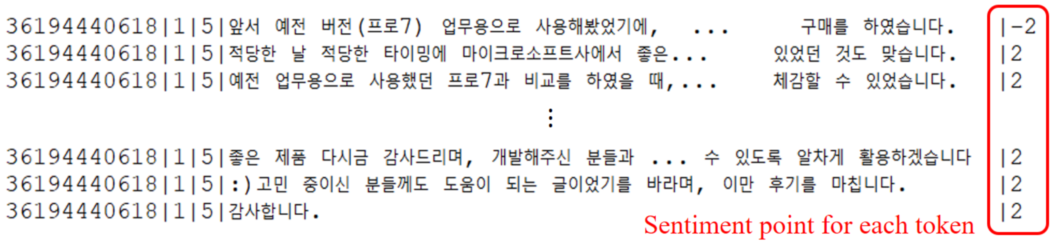
FIGURE 6. Word Cloud for the VoC analysis

From the visualized graph and data from VoC, we can read and recognize the customer's thoughts and mind easily. The Word Cloud graphic is a visual representation that supplements a section of text to help readers better understand an idea or approach a subject from a different angle. If we draw many Word Clouds serially, they will show off trends for the interested problem.

3.4. **Sentiment analysis.** Sentiment analysis methods in text mining are divided into two types:

- How to use the emotional dictionary;
- Machine learning through machine learning.

AI can judge positive/negative through machine learning, but it takes a lot of learning to determine the degree of positive or negative. In this study, an emotional dictionary (dilab.kunsan.ac.kr) is used. (−2: Very negative, −1: Negative, 0: Neutral or undecided, 1: Positive, 2: Very positive.) The result of sentiment analysis is shown in Figure 7.



```
36194440618|1|5|앞서 예전 버전(프로7) 업무용으로 사용해봤었기에,   ...      구매를 하였습니다.   |-2
36194440618|1|5|적당한 날 적당한 타이밍에 마이크로소프트사에서 좋은...      있었던 것도 맞습니다.   |2
36194440618|1|5|예전 업무용으로 사용했던 프로7과 비교를 하였을 때,...      체감할 수 있었습니다.   |2
                                          ⋮
36194440618|1|5|좋은 제품 다시금 감사드리며, 개발해주신 분들과 ... 수 있도록 알차게 활용하겠습니다   |2
36194440618|1|5|:)고민 중이신 분들께도 도움이 되는 글이었기를 바라며, 이만 후기를 마칩니다.   |2
36194440618|1|5|감사합니다.                                                                  |2
```

FIGURE 7. Result of sentiment analysis

4. **Conclusions.** This study proposes a framework for Web crawling and analysis for a product review in e-commerce. An actual Web page review in top portal site is adopted and analyzed. As the related process and task can be connected, many tedious and time consuming tasks are connected automatically leading to autonomous process. The connected whole process can be generated by the concept of RPA process.

In the near future, using chat bots and robots, simple and repetitive tasks will be transformed from human labor to digital labor.

Future research areas are to build a generic framework to utilize RPA concept in various tasks using existing available software, and to create and report many use cases that helps human from tedious and repetitive tasks.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) by the Ministry of Education (NRF-2021R1F1A1056189).

REFERENCES

[1] W. M. P. van der Aalst, M. Bichler and A. Heinz, Robotic process automation, *Computer Science, Business & Information Systems Engineering*, vol.60, pp.269-272, DOI: 10.1007/s12599-018-0542-4, 2018.

[2] K. Tathe and S. Sharma, Data collection using web scrapping with robotic process automation, *International Research Journal of Modernization in Engineering Technology and Science*, vol.4, no.6, pp.1-5, 2022.

[3] R. RK and K. S. Rao, Automated data collection of the examination results in affiliated collegesusing web scrapping techniques, *Dogo Rangsang Research Journal*, vol.12, no.12, 2022.

[4] G. Barbera, L. Araujo and S. Fernandes, The value of web data scraping: An application to TripAdvisor, *Big Data and Cognitive Computing*, vol.121, pp.1-12, 2023.

[5] S. Mehak, R. Zafar, S. Aslam and S. M. Bhatti, Exploiting filtering approach with web scrapping for smart online shopping : Penny wise: A wise tool for online shopping, *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Sukkur, Pakistan, pp.1-5, DOI: 10.1109/ICOMET.2019.8673399, 2019.

[6] M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles and M. Gori, Focused crawling using context graphs, *Proc. of the 26th VLDB Conference*, Cairo, Egypt, pp.527-534, 2000.

[7] M. Lacity and L. Willcocks, Robotic process automation: The next transformation lever for shared services, *The Outsourcing Unit Working Research Paper Series*, pp.1-35, 2016.

[8] K. K Thomas, S. P Anil, E. Kuriakose and N. George, , *International Journal of Information Systems and Computer Sciences*, vol.8, no.2, pp.147-151, 2019.