한국정보전자통신기술학회논문지(jkiiect)'24-10, Vol.17 No.5
논문 24-17-05-323 http://dx.doi.org/10.17661/jkiiect.2024.17.5.323

# Analysis and Utilization of Housing Information based on Open API and Web Scraping

Shin-Hyeong Choi*

# 오픈API와 웹스크래핑에 기반한 주택정보 분석 및 활용방안

최신형*

**Abstract** In an era of low interest rates around the world, interest in real estate has increased. We can collect real estate information using the Internet, but it takes a lot of time to find. In this paper, real estate information from January 2015 to April 2024 is collected from three places to help users more easily collect real estate information of interest and use it for sales. First, by analyzing HTML documents using web scraping techniques, information on real estate of interest is automatically extracted from the website of the platform company. Second, the actual transaction price of the real estate is additionally collected through the open API provided by the Ministry of Land, Infrastructure and Transport. Third, real estate-related news is provided so that users can learn about the future value and prospects of real estate. The simulation results for the data collected in this study show that the lowest price predicted by the ARIMA model is expected to be in May 2024 among the next eight months. Therefore, by following this procedure, real estate buyers can make more efficient home sales by referring to related information including the predicted transaction price.

**요 약** 전 세계적으로 저금리 시대에 부동산에 대한 관심이 증가했으며, 인터넷을 사용하여 부동산 정보를 수집할 수 있지만 찾는 데 많은 시간이 소요된다. 본 논문에서는 2015년 1월부터 2024년 4월까지의 부동산 정보를 수집하여 사용자에게 제공함으로써 관심 있는 부동산 정보를 이용하여 매매에 사용할 수 있도록 돕는다. 첫째, 웹 스크래핑 기법을 사용하여 HTML 문서를 분석하여 플랫폼 기업의 웹사이트에서 관심 부동산 정보를 자동으로 추출하고, 둘째, 국토교통부에서 제공하는 오픈 API를 통해 해당 부동산의 실제 거래 가격을 추가 수집한다. 셋째, 부동산 관련 뉴스를 제공하여 사용자가 부동산의 미래 가치와 전망을 알 수 있도록 한다. 본 연구에서 수집한 데이터에 대해 시뮬레이션한 결과 ARIMA 모델로 예측한 매매가격에 의하면 향후 8개월 중에서는 2024년 5월이 가장 낮은 가격임을 예상할 수 있다. 따라서 이와 같은 절차에 따르면 부동산 매수자는 예측된 거래 가격을 포함한 관련 정보를 참고하여 보다 효율적인 주택 매매를 할 수 있다.

**Key Words** : ARIMA model, Housing Information, Open API, Website, Web Scraping

## 1. Introduction

Big data analysis and artificial intelligence are core technologies of the 4th Industrial Revolution and are becoming the subject of many people's interest and research. Big data companies process and utilize information in an advanced way by analyzing accumulated data, and AI-based development companies conduct research to improve the accuracy of data processing algorithms through machine learning

*Corresponding Author : Division of Electrical, Control & Instrumentation Engineering, Kangwon National University (cshinh@kangwon.ac.kr)
Received September 23, 2024 Revised October 09, 2024 Accepted October 14, 2024

based on large amounts of data. is in progress. Therefore, securing data becomes an essential task for all companies. Meanwhile, with the introduction of generative artificial intelligence technology, related disputes are increasing. In particular, after the advent of ChatGPT, copyright lawsuits over AI-related data and content continue to occur[1]. It is common for smaller companies or individuals to use data on the Internet due to a lack of their own data. As in the case of big data companies or AI-based development companies, individuals collect various data through blogs or information-providing websites on the Internet to solve daily tasks or problems. As interest in housing increases along with stocks, it is essential to collect more accurate information in these areas. In recent years, with the global era of low interest rates, there has been a surge in housing prices, leading to increased interest in real estate both as a dream of homeownership and as an investment. Unlike existing studies that only show real estate price information, this study collects information in various ways, such as past price information, recent transaction trends, and related news about houses of interest, and suggests a method for analyzing and utilizing the collected information.

## 2. Related Works

In the past, owning a home was life's biggest dream, but the appeal of real estate as a means of investing has become even greater now than in the past. According to this trend, real estate investment can bring large profits like stock investment, but the risk burden also increases. More accurate information is essential not only to realize the dream of owning a home, but also to invest in real estate, which has emerged as a means of investing. In the past, to find real estate information, people typically visited real estate agents located in the area of interest and make a decision after hearing information such as rental or sale price, nearby amenities, educational facilities, etc., but recently, through platforms such as popular real estate websites and apps, people can access detailed information about properties available for purchase or rent, including pricing and local area information. In addition, information on housing trends, support programs for purchasing a home, and related taxes is collected through public data portals provided in real time by the government or local governments, and generations prefer communication and information exchange through community forums and social media. In particular, people who are considering real estate as an investment rather than a place of residence attend real estate events, seminars, or conferences and refer to market trend information predicted by experts regarding real estate-related area information and development plans for the area. However, since it is difficult for people with jobs to freely collect information or visit real estate during working hours, most have no choice but to study real estate or collect information on real estate of interest after work. Unlike in the past, it is easier to collect data of interest using the Internet. However, because there is such a wide range of data on the Internet, it takes a lot of time to immediately find the data f interest.

### 2.1. Web scraping

To solve the difficulties and tedious problems of searching mentioned above, the web scraping

technique was developed, and by using it, various types of data can be obtained more easily from websites. In other words, web scraping is to collect necessary data from websites on the Internet and save them in structured formats such as CSV(Comma-Seperated Values) and JSON(JavaScript Object Notation), and it refers to all tasks of collecting data from the Web in an automated manner through a program using data existing on the Internet[2-4].

### 2.2. Open API

In general, an open API (Application Programming Interface) is an application programming interface that can be used publicly. It is also called a public API. It is posted on a web page so that anyone can use it freely, and service owners connected to the Internet can provide access. In other words, it has disclosed a method of accessing through the Internet to provide various information produced or collected in the form of a web service. Open API is a convenient service that can be freely used from outside by requesting information provided by the website and receiving a response using a service key obtained through the proper application process[5,6].

## 3. System design

The purpose of this system is to aggregate housing information in areas of interest to users and provide a time to purchase in the future and an appropriate purchase price. The configuration diagram of this system is shown in Figure 1.

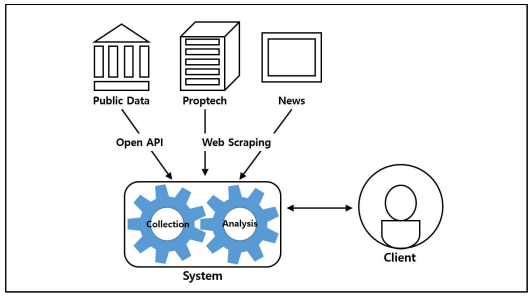

Fig. 1. System diagram

### 3.1. Data collection method

To collect housing information, people can use the services of platform companies that provide real estate information. Housing information is collected by accessing a platform company's website, and the amount or value of housing information that can be collected may vary depending on the policy or service method of the website. In this paper, data collection occurs primarily from three sources. First, real estate information provided by platform companies is collected through web scraping techniques, and second, the actual trading price of the relevant real estate is additionally collected using open API provided by the Ministry of Land, Infrastructure and Transport. Third, it provides news related to real estate so that users can learn about the future value and prospects of the real estate. In this paper, in order to automatically extract real estate information of interest from the website of a platform company that provides real estate information, an HTTP GET request is sent by constructing a URL according to the necessary search conditions, and when the site responds, the scraper analyzes the HTML document and extracts only data with a specific pattern.

Table 1. Real estate information collection code

```
Function get_real_estate_data:
Set base_url to "https://search.naver.com/search.naver"
Set params with parameters
Send a GET request to base_url with params
Receive response
Parse the response text using BeautifulSoup
Extract real estate data based on specified class
Return real estate data
```

Table 2. Actual trading price collection code

```
Function get_apartment_transaction_data:
Set api_url to
"http://apis.data.go.kr/AptTradeInfoService/getTradeList"
Set params with parameters
Send a GET request to api_url with params
Receive response
Extract JSON data from the response
Return data
```

The code that extracts real estate information from platform companies using web scraping techniques is described in Table 1. The requests module is used to retrieve the desired HTML information by sending an HTTP request using the GET method from a website, and the BeautifulSoup module, a parsing library, is used to extract data from the value received through requests.

After checking the listing information of the real estate of interest using the code of Table 1, actual trading price information for the relevant real estate is collected through the open API provided by the Ministry of Land, Infrastructure and Transport. To do this, we call the open API using a service key for the open API obtained in advance, then send an API request by setting the area code and inquiry period for the properties of interest, parse the received response, and add it to the DataFrame, and then save it as a CSV file for further use in predicting trading price information. The code that collects information on the actual trading price of the property through open API is described in Table 2. Based on the API document, an HTTP request is sent based on the URL and parameters set along with the authentication key, and the necessary information is extracted by parsing the JSON data received from the API.

The third is the process of collecting news related to real estate so that users can learn about the future value and prospects of the real estate. In this step, news articles related to real estate collected in steps 1 and 2 are collected using the news search API of a platform company that provides Internet news. The code that sends an API request by setting the relevant real estate search term and period and then parses the received response to extract related news articles is described in Table 3.

Table 3. Real estate-related news collection code

```
Function get_news_data:
Set api_url to
"https://openapi.naver.com/v1/search/news.json"
Set headers with parameters
Send a GET request to api_url with headers and
params
Receive response
Extract JSON data from the response
Return data
```

### 3.2. Data analysis method

Based on the information collected in Section 3.1, we analyze the price trends and consider the future development potential to assess the appropriate timing and fair price for purchasing or selling the property. In general, when analyzing time series data, models such as AR (Autoregressive), MA (Moving average), ARMA (Autoregressive Moving average), and ARIMA (Autoregressive Integrated Moving average)

models are used. The ARIMA model is used to provide regularity to irregular time series data, such as the multi-year real estate trading price changes collected in this study, to analyze price trends[7-10]. In other words, in the ARIMA model, better predictions can be made in non-stationary situations through differencing. Here, p represents the order of the AR part of model, q represents the order of the MA part of the model, and d represents the degree of differencing. The ARIMA model is denoted as ARIMA(p,d,q).

Table 4. Trading price prediction and visualization code

```
Set apartment_name to "A apartment"
Extract apartment data for "A apartment" from
dataframe df
Train ARIMA model on 'Trade_price' with order=(5,1,0)
Forecast transaction prices for May to December 2024
(8 months)
Calculate yearly average transaction prices
Visualize the forecasted and average transaction prices
```

The code that provides visualized data along with monthly trading prices predicted through learning the ARIMA model provided as a Python library for price trend analysis is described in Table 4. In the ARIMA model, the order of differencing represents the number of times differencing is needed to achieve stationarity in the time series data. While it is common to set the order of differencing to 2, in this study, the order of differencing is set to 5 due to the presence of strong seasonality or trends observed in the apartment trading price data used.

## 4. System implementation

This chapter explains the system by dividing it into three parts. The first describes the system environment, and the second describes the data collection process. The third shows the system results.

### 4.1. System environment

The system development environment is a host PC consisting of an Intel(R) Core(TM) i7-3770 CPU and 8GB of RAM, the operating system is Windows 10, and the development language is Python 3.7.9 and Jupyter notebook.

### 4.2. Data collection process

In this paper, real estate information from January 2015 to April 2024 is largely collected from three sources. First, real estate information provided by platform companies is collected through web scraping techniques. Second, the actual sale price of the relevant real estate is additionally collected using the open API provided by the Ministry of Land, Infrastructure and Transport. Third, it provides news related to real estate so that users can learn about the future value and prospects of the real estate.

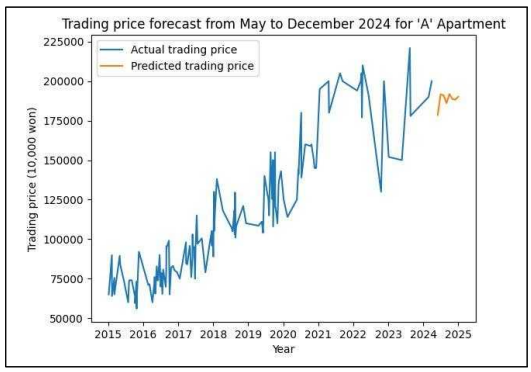


Fig. 2. Apartment trading price trend and forecast

### 4.3. Data analysis

After saving these data as a CSV file, the trading price trends by year and the ARIMA

model predict apartment trading prices after May 2024. Figure 2 is a graph showing this. We can expect that May 2024 will be the lowest price among the next eight months.

## 5. Conclusion

Real estate information for purchasing a house must be up-to-date and accurate, and collecting this information quickly has great economic benefits. People use the Internet to collect real estate information in real time, but because there is so much data on the Internet, it is difficult and time-consuming to find the desired real estate information, and since this information is collected from blogs and SNS, it may contain personal or false information, which may actually increase people's confusion. Therefore, in this paper, we propose a method to provide reliable information to people interested in real estate information in various ways. To this end, we collect the necessary real estate information from reliable real estate information websites and public data portals using web scraping technology and open APIs, and collect real estate-related news along with visualized information including transaction prices predicted by the ARIMA model, so that people can collect housing information more efficiently while considering the future value and prospects of real estate. In the future, we plan to study ways to reflect major economic policies or economic indicators in addition to collecting real estate prices and news information.

## REFERENCES

[1] TheStreet Technology [Website]. (2024. Feb 29) Retrieved from https://www.thestreet.com/technology/copyright-lawsuits-against-openai-microsoft-chatgpt

[2] V. Singrodia, A. Mitra, S. Paul, "A Review on Web Scrapping and its Applications", *2019 International Conference on Computer Communication and Informatics,* pp.1-6, Jan, 2019.

[3] P. Matta, N. Sharma, D. Sharma, B. Pant, "Web Scraping: Applications and Scraping Tools ", *International Journal of Advanced Trends in Computer Science and Engineering, 9(5),* pp.8202-06, Oct, 2020.

[4] C. Lotfi, S. Srinivasan, M. Ertz, I. Latrous, "Web Scraping Techniques and Applications: A Literature Review", *SCRS Conference Proceedings on Intelligent Systems,* pp.381-394, Apr, 2022.

[5] Wikipedia [Website]. (2024 March 5) Retrieved from https://en.wikipedia.org/wiki/Open_API

[6] TechTarget [Website]. (2021 October). Retrieved from https://www.techtarget.com/searchapparchitecture/definition/open-API-public-API

[7] M. Soheila, R. Mohammad. “A Brief Survey on Event Prediction Methods in Time Series”. *Advances in Intelligent Systems and Computing.* pp. 235-246, Jan, 2015.

[8] M. Soheila, R. Mohammad. “Time series forecasting using improved ARIMA”. *2016 Artificial Intelligence and Robotics.* pp. 92-97, Apr, 2016.

[9] L. Yizhuo, S. Weijia. “ARIMA Time Series Modeling and Forecasting of Enterprise Electric Energy Consumption”. *International Conference on Frontiers of Electronics, Information and Computation Technologies.* pp. 497-500, May, 2023.

[10] N. Khin, Y. Yi. “Time Series Data Forecasting System for Stock using TA and ARIMA Model”. *2023 IEEE Conference on Computer Applications.* pp. 72-76, Feb, 2023.

## Author Biography

**Shin-Hyeong Choi** **[정회원]**

- Feb. 2002 : Kyungnam Univ., Computer Engineering, PhD
- Sept. 2003 ~ current : Kangwon National Univ., Div. of Electrical, Control & Instrumentation Engineering, Professor

〈Research Interests〉 Embedded system, IoT, Machine learning