Forecasting Housing Demand with Big Data

Han Been Kim[1], Seong Do Kim[2], Su Jin Song[3], Do Hyoung Shin[4*]

Abstract: Housing price is a key indicator of housing demand. Actual Transaction Price Index of Apartment (ATPIA) released by Korea Appraisal Board is useful to understand the current level of housing price, but it does not forecast future prices. Big data such as the frequency of internet search queries is more accessible and faster than ever. Forecasting future housing demand through big data will be very helpful in housing market. The objective of this study is to develop a forecasting model of ATPIA as a part of forecasting housing demand. For forecasting, a concept of time shift was applied in the model. As a result, the forecasting model with the time shift of 5 months shows the highest coefficient of determination, thus selected as the optimal model. The mean error rate is 2.95% which is a quite promising result.

Keywords: Big data, Housing demand, Housing prices, Forecasting model, Search queries, Regression

Ⅰ. INTRODUCTION

Housing market is hard to change immediately in supply according to the changes in demand. The difficulty in maintaining the balance between supply and demand in housing market indicates that forecasting housing demand is very important. Housing price is a key indicator of housing demand. However, it is very difficult to forecast housing price because it is very sensitive to the fluctuation of social and economic environments.

Korea Appraisal Board has released Actual Transaction Price Index of Apartment (ATPIA) monthly since January 2006 through their real estate statistics information system called R-ONE. ATPIA is a price index based on the data of actual transaction prices which are required to report to the government. ATPIA has been collected and released since January 2006. The index presents the relative changes of apartment price compared to the price at the reference point of January 2006 of which index is set as 100 [1].

ATPIA is a statistical survey using the reported transaction data, so it is useful to understand the current level of housing price. However, there can be a time lag to collect data because a transaction price can be reported within 60 days from the transaction date. Moreover, ATPIA is limited to present the current price, and it does not forecast future prices.

The limitations of ATPIA can be addressed by using big data. Big data-based forecasting is one of the new forecasting analysis techniques gaining worldwide attention. Big data enables to find out hidden factors for forecasting by utilizing mass data and to analysis much faster than conventional statistical analysis. It is greatly helpful for forecasting [2].

With the emergence of smart phones, the frequency of internet search queries is increasing exponentially. The search frequency regarding housing also has significantly increased. Public interests in housing market and financial conditions affect housing demand and price. Therefore, if their trends are caught through the change in internet queries, it is possible to forecast housing prices through big data such as the frequency of internet search queries. The objective of this study is to develop a forecasting model of ATPIA using the frequency data of internet search queries, as a part of forecasting housing demand..

Ⅱ. BIG DATA ANALYSIS WITH INTERNET QUERIES

The type of big data for forecasting analysis is very diverse. Particularly, studies for forecasting based on internet search queries has been carried out in a variety of industries due to the explosive increase of search volume. Google developed the Google Flu Trends using internet search queries to alert influenza [3]. Goel et al. [4] claimed that big data-based approaches using internet search queries are useful in forecasting movie, video game, and music sales when it is hard to collect and access to data for traditional forecasting approaches. Kim and Shin [5] developed an automatic handling system for internet search query data. Significant reduction in time to handle the data by using the system indicates the possibility of easier adoption of big data in forecasting urban infrastructure demand.

There are some studies on the correlation between internet search queries and housing market. For example, McLaren [2] presented that predictability of forecasting model of housing market can be increased by adding large amounts of internet search data to the baseline from the regression model for housing index in UK. Song et al. [6] claimed the potential of internet search queries as the forecasting factor for housing market, by analyzing correlation between frequency of internet search queries and home sales index Kim and Yu [7] presented that internet search queries have a causal effect on housing price and housing price also has a causal effect on trading volume.

As shown above, the frequency of internet search queries can be useful for forecasting. This study adopted internet search queries as an independent factor for a forecasting model of ATPIA. To prevent the accidental omission of important but unsuspected queries, this study

[1] Graduate Student, Civil Engineering, College of Engineering, INHA UNIVERSITY, beenkkk@inha.edu
[2] Graduate Student, Civil Engineering, College of Engineering, INHA UNIVERSITY, frozen0327@inha.edu
[3] Graduate Student, Civil Engineering, College of Engineering, INHA UNIVERSITY, ssconsys@inhaian.net
[4*] Associate Professor, Corresponding Author, Civil Engineering, College of Engineering, INHA UNIVERSITY, 100, Inha-ro, Nam-gu, Incheon 402-751, Korea, dhshin@inha.ac.kr

tried to include internet search queries as many as possible.

Ⅲ. FORECASTING MODEL

A. Data Collection and Handling

NAVER and Google have been leading search engines in Korea. They release the search frequency data formed as relative frequency through NAVER Trend and Google Trends, respectively. [8, 9] According to Korea search market share, NAVER takes 80.24% while Google takes 4.75% [10]. Because NAVER is quite dominant in Korea, it was determined to use NAVER Trend to identify the frequency of internet search queries in this study.

In the meantime, with exponential increasing usage of smart phones and tablets, the volume of mobile internet search data is also significant and cannot be ignored compared to one of PC internet search data. Therefore, it was determined to use both of PC and mobile internet search data in this study. However, there is a technical problem with integrating PC and mobile internet search data. NAVER Trend provides relative frequency of internet search queries compared to one of a reference query or point. In addition, NAVER Trend releases these data for PC and mobile internet searches separately. Due to the lack of the absolute frequency, it is hard to combine PC and mobile internet search queries as they have different reference values.

To resolve this problem, NAVER Search AD was employed. NAVER Search AD releases absolute frequencies of limited specific queries in both PC and mobile searches [11]. Using the absolute frequencies of those bridge queries from NAVER Search AD as connection points between PC and mobile search queries, the absolute frequencies of each of other queries in PC and mobile searches were estimated and combined. Figure 2 shows an example of a pattern of absolute search frequency with a query of 'civil' from NAVER Search AD.

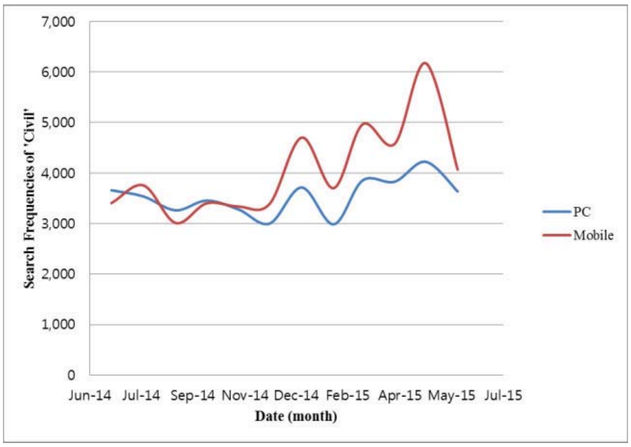

FIGURE 1 Example of search frequency from NAVER Search AD

The key of big data analysis is to utilize messy but massive data [12]. Such approach enables to get macro-level insights and to prevent omitting unsuspected important variables. In this study, the candidate search queries, among which significant queries are going to be selected and used as independent variables, were made up with approximately 3,400 queries. The candidate queries came from the queries cautiously selected from 'NAVER Trends Yearbook 2009' and the queries related housing and economy derived from a brainstorming.
The frequency data of internet search queries released by NAVER is on weekly basis. The data in PC searches is available from January 2007 while the data in mobile searches is available from in late June 2010. This study employed data both in PC and mobile searches from July 2010.

Another thing to consider is the internet traffic. The volume of the internet usage has changed continuously, sometime very dramatically. However, the data released from NAVER does not consider changes in the internet traffic. If the frequency of a certain query at one time point is equivalent to one at another time point while the internet traffics are different between the two time points, the proportions of the query among the total traffics are different. To reflect the fluctuation of the internet search traffics, the internet traffic index was made by dividing the sum of frequencies at each point by the average of the summed frequencies in this study. Then, the frequency of each query at each point was adjusted by dividing it by the corresponding traffic index. Figure 3 shows the pattern of the internet search traffics derived from the sum of query frequencies at each point.

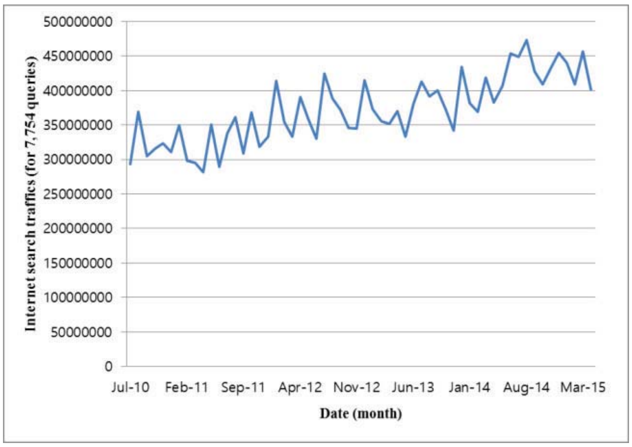

FIGURE 3 Internet search traffics

B. Model description

To select search queries for the forecasting model among the candidate search queries, a correlation between the adjusted frequency of each query and ATPIA was analyzed and ranked through the about 3,400 candidate search queries. In this process, a concept of time shift between the independent variable (frequency) and the dependent variable (ATPIA) was applied for forecasting. As shown in Figure 4, if the independent variable and dependent variable occur at the same time, the model does not provide forecasting. To make a model with forecasting, this study used the independent variable occurring prior to the corresponding dependent variable as shown in Figure 4. The correlations calculated with the time shift of 0 month through 11 months by 1 month.

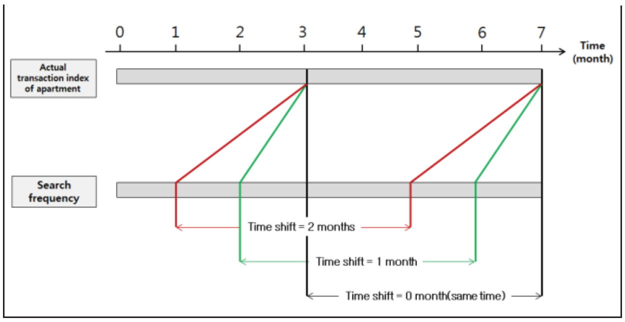

FIGURE 4 Concept of Time Shift for forecasting

The absolute values of the correlation coefficients are used as the correlation between the frequency of each search queries and ATPIA. Tables 1 and 2 show examples of the top 10 queries with high correlation with time shift of 0 month and 1 month, respectively.

TABLE 1 Top 10 queries of high Correlation (Time Shift=0month)

Rank	Query	Correlation Coefficient	Correlation
1	Tuna Kimchi stew	0.8882	0.8882
2	Consumers Union of Korea	-0.8862	0.8862
3	Pathfinding	0.8828	0.8828
4	Optimus	-0.8643	0.8643
5	Timer	0.8629	0.8629
6	LED Lighting	0.8402	0.8402
7	Calculator	0.8390	0.8390
8	Shoes	0.8365	0.8365
9	Bluetooth Earphones	0.8270	0.8270
10	YAMAHA	0.8265	0.8265

TABLE 2 Top 10 queries of high Correlation (Time Shift=1month)

Rank	Query	Correlation Coefficient	Correlation
1	Pathfinding	0.8721	0.8721
2	Namsan Cable Car	0.8597	0.8597
3	Tuna Kimchi stew	0.8511	0.8511
4	Calculator	0.8496	0.8496
5	Timer	0.8440	0.8440
6	Consumers Union of Korea	-0.8398	0.8398
7	LED Lighting	0.8359	0.8359
8	Stir-fried anchovies	0.8355	0.8355
9	Café	0.8246	0.8246
10	Shoes	0.8238	0.8238

After making the rank of search queries with high correlation for each time shift, the values of the independent variable were made by adding the frequencies of search queries from the highest correlation in sequence. For each value of the independent variable from the summation, a forecasting model with each time shift was created and tested. In this manner, the candidate forecasting models were made from regression analysis between the independent variable (sum of frequencies) and the dependent variable (ATPIA). In addition, K-fold (10-folds in this study) cross validation was conducted to verify the candidate models. For each time shift, the model with the highest K-fold average coefficient of determination was selected. Figure 5 and Table 3 show the highest K-fold average coefficient of determination for each time shift and the corresponding number of queries included in the independent variable.

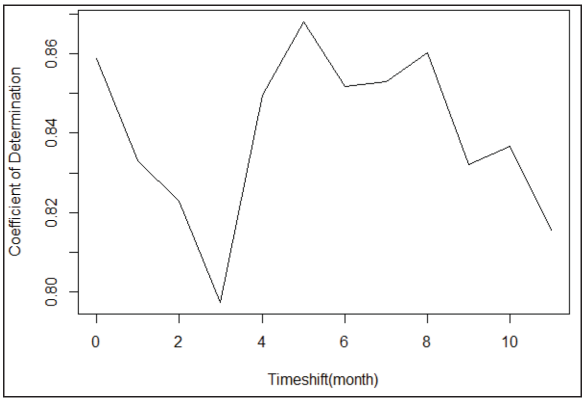

FIGURE 5 K-fold average coefficient of determination of the model in accordance with the Time Shift

TABLE 3 K-fold average coefficient of determination of the with the Time Shift

Time shift (month)	Number of queries included in the model	K-fold average coefficient of determination
0	35	0.8589
1	34	0.8331
2	29	0.8228
3	39	0.7974
4	10	0.8496
5	7	0.8681
6	7	0.8518
7	7	0.8530
8	5	0.8602
9	16	0.8321
10	34	0.8367
11	3	0.8157

As shown in Figure 5 and Table 3, the optimal time shift value with the maximum K-fold average coefficient of determination is 5 months and the corresponding independent variable is the sum of the 7 search queries. Thus, the forecasting model with the time shift 5 months and 7 search queries was selected as the final forecasting model. Table 4 shows that the list of search queries (key queries) included in the independent variable of the final forecasting model.

TABLE 4 Key queries (Time Shift=5months)

Query			
Calculator	Horse Racing	77 Size	Glamour
Stir-fried anchovies	KOSDAQ		Radio

C. *Results*

The final forecasting model was derived from the process above as following:

$$Y = 2774.5263X + 133.8068$$

Y: Forecasted ATPIA
X: Sum of the frequencies of key queries
(with time shift of 5 months)

Table 5 shows that the statistics of final forecasting model. The model and its coefficients were revealed significant.

TABLE 5 Statistics of the final forecasting model

	Estimate	Std. Error	t value	P value
Intercept	133.8068	0.5497	243.40	2e-16
Variable	2774.5263	174.3938	15.91	2e-16
Multiple R-squared	0.8695	Adjusted R-squared		0.866
F-statistic (DF: 1, 38)	253.1	P value		2.2e-16

Figure 6 and Table 6 show the comparison results between the estimated ATPIAs from the forecasting model and the actual ATPIAs which were unknown yet when the model was developed. The black solid line represents the actual ATPIA from May 2011 to August 2014 and the red dotted line describes the corresponding estimated ATPIA with the sum of key queries from December 2010 to March 2014 (time shift of 5 months).

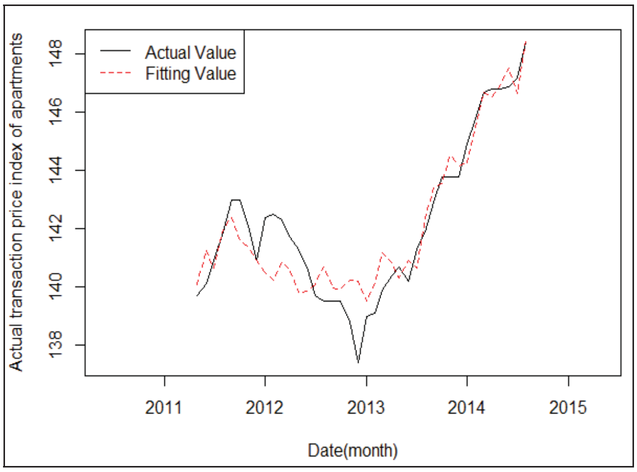

FIGURE 6 Comparison of the estimated ATPIA versus actual ATPIA

As shown in Figure 6, the estimated ATPIA from the forecasting model is similar to the actual ATPIA. The forecasting model also fits the data well as mentioned above that the adjusted coefficient of determination of the model is 0.866.

As shown in Table 6, the error rates of the forecasted ATPIA versus the actual ATPIA range from 1.51% to 3.89% and mean error rate is 2.95% which is a quite promising result.

TABLE 6 Error rate of the forecasted ATPIA versus the actual ATPIA

Mon	Actual value	Forecasting value	Error rate
9	149.5	147.237	1.51
10	150.4	146.5832	2.54
11	151	146.9788	2.66
12	151.5	146.875	3.05
1	152.4	147.0486	3.51
2	153.5	148.1152	3.51
3	155	148.9733	3.89

Ⅳ. CONCLUSIONS

This study developed a forecasting model of ATPIA using the frequency data of internet search queries, as a part of forecasting housing demand. By using internet search query data, the approach presented in the study does not require high cost or time lag for data collection. Due to the easy access to the data, the forecasting model can be quickly updated. The validation results of the forecasting model present the mean error rate of 2.95% which is a quite promising result.

Some challenges were recognized in developing the forecasting model. First, there is unavoidable accumulation of errors in data handling process. NAVER Trend releases only the relative frequency of internet search queries in an integral number format. Due to the lack of decimal values, the inevitable errors could occur. This lack indicates the necessity of further study on conversion of the relative frequencies to absolute ones.

Second problem is the existing limit in setting the candidate search queries. The candidate queries in this study were set mainly based on 'NAVER Trends Yearbook 2009' which is the only released one. Thus the candidate queries may reflect overly the phase of the times of Korea in 2009. The search queries reflecting those days can be a noise when developing the model. This problem can be improved by using various data collection techniques.

The approach presented in this study can be adopted in forecasting demand of other infrastructure such as subways, expressways, airports, etc. This study shows the potential of big data in forecasting infrastructure demand. It also indicates the necessity of developing big data techniques and methodologies tailored to forecasting infrastructure demand.

ACKNOWLEDGEMENTS

This work was supported by INHA UNIVERSITY Research Grant.

REFERENCES

[1] Korea appraisal board, "R-ONE, Real Estate Official Statistics Network", http://www.r-one.co.kr/rone/, accessed 2015
[2] McLaren, N., "Using Internet search data as economic indicators", Bank of England Quarterly Bulletin, 2011

[3] Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L., “Detecting influenza epidemics using search engine query data”, Nature, Vol 457.7232, pp.1012~1014, 2009

[4] Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J., “Predicting consumer behavior with Web search”, National academy of sciences, Proceedings of the National academy of sciences, 107(41), pp.17486-17490, 2010

[5] Kim, S. and Shin, D., “Development of an Automatic Database Handling System of Internet Queries for Big Data-based Demand Forecast of Urban Infrastructure”, Korean Society of Civil Engineers(KSCE), Proceedings of KSCE Annual conference, pp.1129-1130, 2014

[6] Song, S., Kim, S. and Shin, D., “Urban Infrastructure Planning of Big Data Era”, Korea Institute of Construction Engineering and Management(KICEM), Proceedings of KSCE Annual conference, 2014-11, pp.13-17, 2014

[7] Kim, D. and Yu, J., “A Dynamic Relationship Between Internet Search Activity, Housing Price, and Trading Volume”, KRERI Research Reports, 24(2), pp.125-140, 2014

[8] NAVER, “NAVER Trend”, http://trend.naver.com, accessed 2015

[9] GOOGLE, “GOOGLE Trends”, https://www.google.com/trends/, accessed 2015

[10] AceCounter Trend Report, “AceCounter”, http://www.acecounter.com/www2/education/trendReportDetail.amz?rno=90&cpage=1&f_serach1=&i_serach= , accessed 2015

[11] NAVER, “NAVER Search AD”, http://searchad.naver.com, accessed 2015

[12] Mayer-Schönberger, V., and Cukier, K., “Big data: A revolution that will transform how we live, work, and think”, Houghton Mifflin Harcourt, 2013