Frameworks of Big Data based Web Data Attribute Analysis for Home Sales Index Prediction*

주택매매지수 예측을 위한 빅 데이터 기반 웹 데이터 속성 분석

Ko, Yongho** · Han, Seungwoo*** · Lee, Sangyoub****
고 용 호 · 한 승 우 · 이 상 엽

Contents

국문초록

1. 내 용

(1) 연구목적

주택매매지수(Home Sales Index : HSI)는 한 국가의 부동산산업과 건설산업의 경제적인 현황을 파악할 수 있는 최적의 지표로서 공공부문의 건설 및 주택정책 수립뿐만 아니라 민간부문의 사업계획 및 투자에 많은 영향을 미친다. 산업사회의 발달과 함께 주택매매지수에 영향을 미치는 변수들이 다양해짐에 따라 데이터의 확보를 위한 다양한 방법론의 제안과 함께 주택매매지수의 예측을 위한 새로운 연구가 지속적으로 이루어지고 있다. 이에 본 연구에서는 정보기술의 발달로 가능해진 새로운 분석방법인 빅데이터 분석에 기반하여 웹데이터 속성을 분석하고 이를 통한 예측모델의 활용성을 제고하고자 한다.

* 이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2015R1D1A1A01058221)
** 주 저 자 : 인하대학교 건축공학과 박사과정, yonghofan@gmail.com
*** 공동저자 : 인하대학교 건축공학과 교수, 공학박사, shan@inha.ac.kr
**** 교신저자 : 건국대학교 부동산학과 교수, 건설경영학박사, sangyoub@konkuk.ac.kr
▷ 접수일(2016년 3월 10일), 수정일(1차 : 2016년 3월 28일), 게재확정일(2016년 5월 20일)

(2) 연구방법

본 연구에서는 기존의 인과관계에 따른 데이터 분석방법론이 아닌 연구방법으로는 우선 주택매매지수 데이터를 구축하고 네이버 포탈에서 검색된 데이터를 기반으로 예측모델을 구성하였으며 이를 상관분석과 다항회귀분석 및 ANOVA분석을 통해 통계적인 유의성을 확인하였다.

(3) 연구결과

제안된 주택매매지수 예측을 위한 빅 데이터 기반 웹데이터 속성분석 연구결과, 상관분석에 따라 데이터 세트를 구분하고 각 세트에 대해 웹데이터를 4개의 부분으로 구분하였다. 이를 통해 다항회귀분석을 실시하여 도출된 R2 및 수정된 R2를 기반으로 최종 데이터세트를 구성하고, MAE, NMSE 확인을 통해 모델의 적정성을 제고하였다.

2. 결 과

본 연구에서는 정확한 주택매매지수의 예측을 위해 인터넷 포탈인 네이버 포탈의 검색어 분석기능에 따라 데이터세트를 분석하고 예측모델로서의 통계적인 유의성을 검토하였다. 이러한 매매지수 예측의 개선을 위한 다양한 모델개발 연구를 통해 향후 부동산 및 건설사업의 기획 및 설계단계에서 매매지수의 패턴을 확인하고 이에 따른 유용한 지표로서 활용가능한 시사점을 제공하고자 한다.

3. 핵심어

• 부동산산업, 주택매매지수, 웹데이터, 지수예측, 검색어, 빅데이터

ABSTRACT

The HSI (Home Sales Index) has been considered one of crucial factors for analyzing economic trends in the construction and real estate industry. It has been investigated that researches of appropriate estimation of HSI have been conducted in the real estate research field generally. However, a precise prediction methodology of HSI has not yet been suggested due to the difficulty of collecting valuable information that has significant influence on the HSI. This study applies a new approach of big data analytic methodology that uses search query data collected from Naver trend. It is expected that the analysis methodology suggested in this paper provides the construction industry with valuable reference that can be used and developed for the establishment of efficient economic strategies and plans in the planning and preliminary design phase of projects.

KEY WORDS : Real Estate Industry, Home Sales Index, Web Data, Index Prediction, Search Query, Big Data

I . Introduction

In Korea, as the residential construction market has played a key role among the business domain of construction industry, lots of prior studies have been performed for housing market such as housing prices, attributes, trading volumes, etc.[1)2)3)] Based on which, this study intends to extend its application to the index representing home sales by implementation of newly proposed methodology. HSI (Home Sales Index) is an index presented by the Kookmin Bank to see the trend of the domestic home sales.[4)] In construction industry, the HSI is frequently used for managing economic strategies and plans in the planning and preliminary design phase of construction projects. Numerous researches have suggested various methodologies for the efficient estimation of HSI.[5)6)7)] However, previous studies have been focusing solely on the estimation process and have neglected the prediction of future HSI values which can provide more valuable information for more efficient strategies development. The process of HSI prediction involves various factors to be considered and also requires massive data collection. The collection of such data for reliable prediction results is a hard task and time consuming.

Researchers have also been conducting studies on finding the appropriate raw data for data analysis based predictions but have shown limitations in suggesting reliable prediction results.[8)9)10)] Therefore, the purpose of this study is to suggest a framework of data collection and analysis for the of the feasible prediction method development. In accordance with the preliminary review of analytic methodologies for HSI assessment based on big-data analysis,[11)] the research

1) Repkine, A., and Song, S., "A Spatial Approach to the Hedonic Pricing of Apartment Attributes", *Korea Real Estate Academy Review,* 2015, vol.63, pp.5-17.

2) Lee, J., and Lee, C., "Analysis on the Determinants of Korea Housing Price Index in Unstable Housing Market by Volatility of the Housing Price", *Korea Real Estate Academy Review,* 2014, vol.59, pp.203-216.

3) Sim, S., "Panel Analysis of Relationship between House Sales Prices and Trading Volume", *Korea Real Estate Academy Review,* 2015, vol.63, pp.18-31.

4) Bang, K., "Real Estate Terms Dictionary", *Buyeonsa.* 2011. pp.1-561.

5) Ahn, J.J., Byun, H.W., Oh, K.J., and Kim, T.Y. "Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting", *Expert Systems with Applications*, 2012, vol.39, no.9, pp.8369-8379.

6) Beracha, E., and Wintoki, M.B., "Forecasting residential real estate price changes from online search activity", *Journal of Real Estate Research,* 2013, vol.35, no.3, pp.283-312.

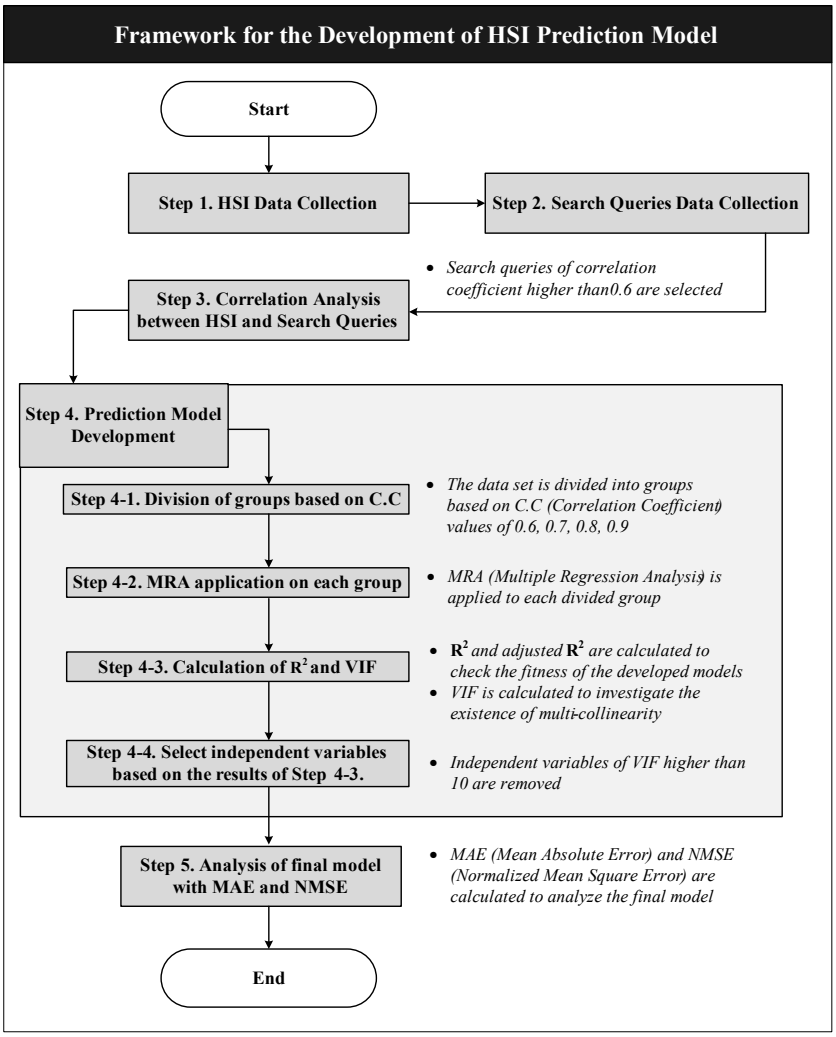
7) Kim, D. and Yu, J., "A Dynamic Relationship Between Internet Search Activity, Housing Price, and Trading Volume", *Korea Real Estate Review,* 2014, vol.24, no.2, pp.125-140.

8) Bollen, J., Mao, H., and Zeng, X., "Twitter mood predicts the stock market," *Journal of Computational Science,* 2011, vol.2, no.1, pp.1-8.

9) Choi, H., and Varian, H., "Predicting the present with Google Trends," *Economic Record,* 2012, vol.88, issue s1, pp.2-9.

10) Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J., "Predicting consumer behavior with Web search," *Proceedings of the National academy of sciences,* 2010, vol.107, no.41, pp.17486-17490.

〈Figure 1〉 HSI prediction model development framework



process for this study is as shown in Figure 1.

First, the HSI data has been collected from a public data sheet by the KB Bank of Korea. Secondly, the independent variables to be used in the prediction model are found, where search query data has been selected from a public database provided by Naver. And, correlation analysis has been applied to deduct statistically significant search queries that can be used for developing a regression model. Finally, the fourth step is divided into four parts. The first part is to divide the search query data into groups based on the C.C

11) Ko, Y., Han, S., and Lee, S., "Preliminary analytic methodologies for home sales index assessment based on big-data analysis", *Review of Real Estate and Urban Studies*, forthcoming 2016, vol.8, no.2. pp1-13.

(Correlation Coefficient). More detail explanation is provided in the below chapters. The second part is to develop a MRA (Multiple Regression Analysis) model for each of the divided groups. R square and VIF (Variation Inflation Factor) values are calculated in the third part. The R square is calculated to investigate the goodness of fit of the developed models. The VIF is calculated to investigate the existence of multicollinearity between variables. In the fourth part, the independent variables for the final model is selected based on the results of part three. Generally, a MRA model is verified through the R square values. However, in this study, an additional investigation on the prediction accuracy has been performed by using MAE (Mean Absolute Error) and NMSE (Normalized Mean Square Error). The calculation process and detail explanation of MAE and NMSE are provided in Chapter V.

II. Literature Review

1. Home Sales Index

From prior studies,[12][13] it is found that a positive correlations exists between housing price and trading volume with various social factors, so more precise assessment and prediction of the HSI are required for better feasible economic planning in construction projects. HSI is generally assessed based on subjective decisions from an empirical, and not a comprehensive, analysis of the facts due to the difficulties and practical limitations by the collection and analysis of actual data, which have a significant influence on the HSI.

2. Web data based Prediction

Recently, Beracha and Wintoki (2013)[14] have conducted a research on forecasting residential real estate price changes based on online search activities. It has been stated that the intention of buying a home is revealed by many potential home buyers when they turn to the web to search for their future residence. In addition, Tuna (2010)[15] has noted that according to Google's chief economist, search queries such as "unemployment office" and "jobs" help predict initial jobless claims. Choi and Varian (2009)[16] have shown that automobile sales and tourism can be forecasted based on web search activities. Da et al. (2011)[17] and Joseph et al. (2011)[18] also have shown that search intensity for stock tickers can predict

12) Beracha, E., and Wintoki, M.B., op. cit. pp.283-312.

13) Kim, D. and Yu, J., op. cit. pp.125-140.

14) Beracha, E., and Wintoki, M.B., op. cit. pp.283-312.

15) Tuna, C., "New Ways to Read Economy," *The Wall Street Journal,* 2010, April 8. (http://www.wsj.com/articles/SB10001424052702303395904575158030776948628)

16) Choi, H., and Varian, H., "Predicting the present with Google Trends", 2009 Technical report, Google, pp.1-20.

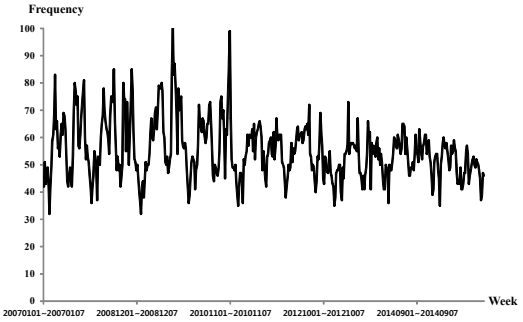
17) Da, Z., Engelberg, J., and Gao, P., "In Search of Attention," *Journal of Finance,* 2011, vol.66, no.5, pp.1461-1499.

future abnormal stock returns and trading volume. A research by Ginsberg (2009) has suggested that predictions of social behavior can be predicted by using search query data as raw data.[19] The methodologies used in the above studies are big data analytic methods that has not been used in the past. As can be found on the prior literature[20], such big data analytic methods have proven their efficiency and thus, applied in this study also. However, for the application of such methodologies appropriate raw data must be collected. Therefore it is required to conduct a study on how to collect such data prior to the application of the analysis methods. Also, it is required to develop a methodology for the appropriate attribution of such data to be used as input data in the analysis method. The studies shown above mostly used data provided by Google which is the most popular search engine in the United States. In this study, a domestic HSI data has been collected. Therefore, this study uses data collected from Naver which is considered the most popular search engine in South Korea. Naver Trend provides the amounts of searches in naver.com. In this study, the process of appropriate data selection by conducting a data attribute analysis has been suggested.

Ⅲ. Data Search for HSI Prediction

This chapter deals with a detail explanation of Step. 1 and 2 that has been shown in Figure 1. Search query data represents the surfer's interest on the web. As written in the previous chapter, a research a Ginsberg (2009)[21] that has been progressed by Google, used such data to predict influenza epidemics faster and more accurate than the existing methods for that purpose. The domestic website "Naver.com" provides similar data to Google that can be easily viewed and downloaded through trend.naver.com [22]. Such methodologies are considered as one of big data analytic methodologies which have been proven their efficiency in various research fields. In this study, the search query data has been thus selected

〈Figure 2〉 Query data of "Home"



18) Joseph, K., Wintoki, M.B., and Zhang, Z., "Forecasting Abnormal Stock Returns and Trading Volume Using Investor Sentiment: Evidence from Online Search," *International Journal of Forecasting*, 2011, vol.27, pp.1116-1127.

19) 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, 2009, vol.457, no.7232, pp.1012-1014.

20) Kim, H., and Lee, S., "Analysis on the Consumer Behavior and Location Characteristics based on Big Data for Department Store-Type Discount Store Development", *Korea Real Estate Academy Review*, 2015, vol.63, pp.172-186.

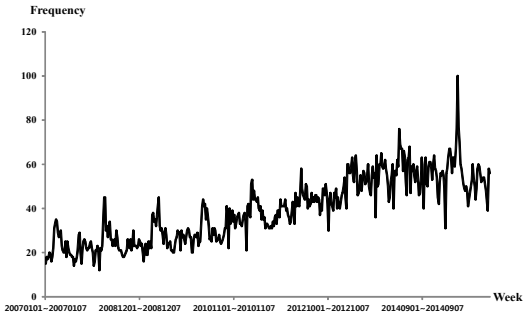
21) Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L., 2009 ibid, pp.1012-1014.

22) Naver Trend, Search Queries available from Internet: http://trend.naver.com. 2016.

to be used as the main source for developing a prediction model focusing on HSI. The search query data provided by Naver Trend is a data showing the amount of searches done per week. The word "Home" for example is shown as Figure 2.

As shown in the figure, the amount of searches is not shown in actual values but in relational values. The maximum value is automatically set to 100 and other values are changed respectively. Search queries are divided into two categories of PC and mobile. The word "Data" for example shows different values when the category is changed from PC to mobile. In this study, the PC version has been selected to be used.

〈Figure 3〉 Query data of "Data" in PC



〈Figure 4〉 Query data of "Data" in mobile

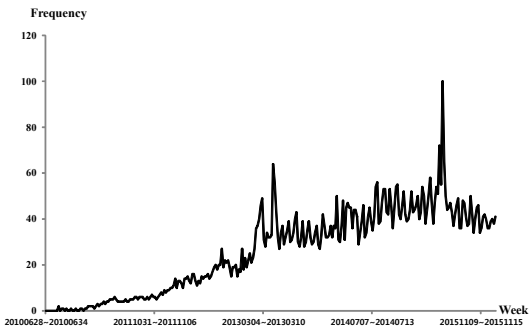


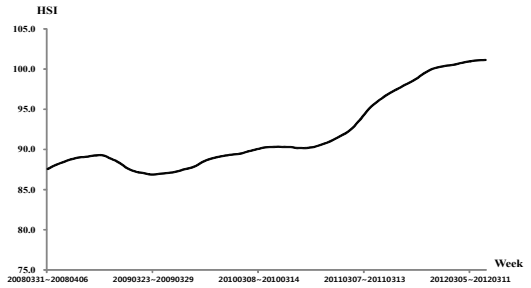
Figure 3 and 4 show each version of data. Naver has been providing search query frequency data since January 1st 2007 in PC version. However, the mobile version has been provided since June 28th 2010 along with the development of the smart phone market. Thus, only the PC version has been used since the collected HSI data starts from March 31st 2008 which exists two years prior to the mobile version data.

In this study, the overall search queries existing in Naver Trend have been set as a population. Six hundred random words related to HSI have been deducted by conducting brainstorming in order to be used as a sample data representing the population data. The deducted words consist not only of words related to HSI but also words used in daily life such as Lotteria, Galleria, Used car and etc. The next chapter deals with the selection of independent variables used for the development of HSI prediction model that can be referred for more detail information.

The HSI data that has been used in this study has been collected from a public data that is provided by the domestic KB Bank. The collected HSI data has a minimum value of 86.9 and a maximum value of 101.1. The mean and standard deviation are 91.9 and 4.666 respectively.

Figure 5 shows the graph of the collected HSI data where the x axis shows time in weeks and y axis the HSI. The HSI data has been collected in a period of March 31st 2008 to April 23rd 2012.

〈Figure 5〉 HSI data



Ⅳ. Prediction Model

1. Data selection

As shown in Step. 3 of Figure 1, this research uses correlation analysis method to select the significant variables that can be used in the analysis. C.C is a value representing the strength of a linear relationship between two data sets. C.C can be calculated based on equation (1).

$$CC = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y} \quad (1)$$

where,

$\bar{x}$ and $\bar{y}$ are the mean values of each data set X and Y

s_x and s_y are the standard deviations of each data set X and Y

It has been studied that a C.C higher than 0.6 shows a strong relationship

〈Table 1〉 Sensitivity analysis of correlation analysis

C.C range	Search Queries
0.6≤|C.C|〈0.7	Ownership, Seoul housing prices, Nonmonetary institutions, Daejeon housing prices, Financial institutions, Marketing ownership, Premium ownership
0.7≤|C.C|〈0.8	Dolce & Gabbana, Officetel, Luxury shoes, Daegu housing prices, Smart TV, National housing bonds, A certified copy of the register, Bank loans, Contract area, Dongtan new city, Consumer finance, Whole certified copy of the register
0.8≤|C.C|〈0.9	CK, Bulgari, Credit reference, Lease contract, Jugong apartment, Latest released movie, Paul Smith, Kwangju housing prices, Minwon 24, Small houses, 4bay, Galleria, Nice auction, Dr. Apart, Restaurants, Yamaha, Joins land, Brokerage commissions, Seal certification for Sale, Pusan housing prices, Newlyweds home, Lotteria, Deposit return, Property views, Ladder truck cost, Proof of income document, Cleanup system
0.9≤|C.C|〈1	IBK Ki-up Bank, ID card authenticity check, Used car, Waste sticker, Apartment ownership

〈Table 2〉 ANOVA table of each model

Model		Sum of Squares	df (degree of freedom)	Mean Square	F	p-value
C.C ≥ 0.6 (51 variables)	Regression	4431.260	51	86.887	193.525	0.000
	Residual	70.040	156	0.449	-	-
	Total	4501.3000	207	-	-	-
C.C ≥ 0.7 (44 variables)	Regression	4424.310	44	100.553	212.887	0.000
	Residual	76.990	163	0.472	-	-
	Total	4501.300	207	-	-	-
C.C ≥ 0.8 (32 variables)	Regression	4416.338	32	138.011	284.267	0.000
	Residual	84.962	175	0.485	-	-
	Total	4501.300	207	-	-	-
C.C ≥ 0.9 (5 variables)	Regression	4259.953	5	709.992	591.302	.000
	Residual	241.346	202	1.201	-	-
	Total	4501.300	207	-	-	-

〈Table 3〉 Regression results (t-test) of C.C part I

By C.C value * ≥0.6 (51 variables) ** ≥0.7 (44 variables)	Unstandardized Coefficients				Stand. Coefficients		t		p-value		-VIF	
	B		Std. Error		Beta							
	*	**	*	**	*	**	*	**	*	**	*	**
(Constant)	99.227	100.286	1.183	1.116			83.893	89.877	.000	.000		
CK	-.013	-.002	.019	.019	-.031	-.004	-.672	-.088	.502	.930	21.504	19.264
Dolce & Gabbana	.004	.003	.011	.010	.009	.006	.387	.285	.699	.776	5.143	4.558
Ownership	.016	-	.005	-	.063	-	3.198	-	.002	-	3.915	-
Bulgari	.050	.057	.020	.020	.108	.124	2.448	2.860	.015	.005	19.503	17.950
Seoul housing prices	-.005	-	.004	-	-.029	-	-1.221	-	.224	-	5.721	-
Credit reference	-.055	-.058	.015	.015	-.132	-.139	-3.592	-3.867	.000	.000	13.490	12.262
Officetel	.000	.021	.021	.021	.000	.032	.015	1.026	.988	.306	10.542	9.314
Nonmonetary institutions	.005	-	.010	-	.012	-	.492	-	.624	-	6.093	-
Lease contract	.001	.000	.014	.014	.004	.000	.077	-.011	.939	.992	21.467	20.908
Jugong apartment	.003	.003	.012	.011	.011	.010	.245	.259	.807	.796	20.840	14.991
Latest released movie	-.025	-.023	.008	.008	-.075	-.070	-3.242	-3.100	.001	.002	5.316	4.871
Luxury shoes	-.022	-.023	.013	.013	-.059	-.062	-1.686	-1.803	.094	.073	12.236	11.232
Paul Smith	-.033	-.034	.013	.013	-.091	-.093	-2.502	-2.710	.013	.007	13.169	11.313
IBK Kiup Bank	.019	.014	.015	.015	.054	.040	1.263	.915	.208	.362	18.655	18.109
Kwangju housing prices	.036	.049	.018	.018	.069	.094	1.958	2.725	.052	.007	12.435	11.381
Daegu housing prices	.009	.001	.013	.013	.015	.001	.672	.052	.503	.959	5.203	4.913
Daejeon housing prices	-.001	-	.008	-	-.003	-	-.142	-	.887	-	5.409	-
Minwon 24	-.006	.004	.018	.017	-.015	.011	-.326	.229	.745	.819	21.459	20.050
Small houses	-.005	-.011	.014	.014	-.012	-.025	-.381	-.806	.704	.421	9.603	9.086
Smart TV	-.012	-.011	.007	.007	-.051	-.047	-1.632	-1.496	.105	.137	9.699	9.605
Financial institutions	-.004	-	.006	-	-.016	-	-.695	-	.488	-	5.125	-
ID card authenticity check	.037	.046	.016	.016	.123	.151	2.341	2.912	.021	.004	27.624	25.675
4 bay	.018	.022	.007	.007	.065	.080	2.594	3.167	.010	.002	6.375	6.132
Galleria	-.038	-.040	.009	.009	-.143	-.149	-4.217	-4.590	.000	.000	11.593	10.013
National housing bonds	-.003	.003	.011	.011	-.010	.010	-.233	.237	.816	.813	19.157	17.015
Nice auction	-.018	-.019	.009	.009	-.060	-.066	-1.939	-2.194	.054	.030	9.522	8.591
Dr Apart	.012	.018	.018	.017	.044	.066	.654	1.027	.514	.306	45.045	39.191
A certified copy of the register	-.007	-.010	.014	.014	-.022	-.034	-.464	-.754	.644	.452	21.675	19.162
Restaurants	-.025	-.030	.008	.008	-.075	-.092	-2.986	-3.779	.003	.000	6.403	5.675
Yamaha	.027	.012	.014	.013	.099	.045	1.965	.940	.051	.348	25.626	22.240
Bank loans	.009	.013	.007	.007	.033	.044	1.318	1.767	.189	.079	6.417	6.035
Joinsland	-.037	-.042	.016	.016	-.131	-.150	-2.325	-2.635	.021	.009	32.050	30.926
Brokerage commissions	.001	.004	.012	.012	.004	.015	.079	.312	.937	.755	23.026	21.036
Used car	-.049	-.058	.015	.015	-.158	-.184	-3.309	-3.897	.001	.000	22.791	21.309
Contract area	.004	.004	.003	.003	.026	.026	1.201	1.178	.232	.240	4.854	4.633
Seal certification for Sale	.002	-.006	.007	.007	.008	-.031	.218	-.848	.828	.397	14.069	12.571
Pusan housing prices	-.015	-.015	.010	.009	-.078	-.074	-1.574	-1.601	.117	.111	24.298	20.549
Waste sticker	-.020	-.027	.010	.010	-.123	-.166	-1.994	-2.789	.048	.006	38.258	33.818
Dongtan newcity	.026	.031	.035	.034	.027	.032	.741	.902	.460	.368	13.224	11.947
Consumer finance	-.008	.032	.068	.063	-.003	.013	-.115	.497	.909	.620	8.098	6.662
Marketing ownership	.001	-	.008	-	.002	-	.115	-	.909	-	4.230	-
Newlyweds home	.149	.137	.067	.067	.047	.043	2.222	2.038	.028	.043	4.552	4.330
Premium ownership	.012	-	.006	-	.039	-	2.101	-	.037	-	3.524	-
Whole certified copy of the register	-.003	.006	.007	.007	-.011	.028	-.356	.975	.722	.331	9.653	7.908
Lotteria	.016	.019	.009	.009	.052	.060	1.841	2.114	.067	.036	8.146	7.749
Deposit return	.022	.021	.009	.009	.092	.086	2.368	2.339	.019	.021	15.058	12.923
Property views	.010	.011	.005	.005	.053	.054	2.283	2.310	.024	.022	5.337	5.237
Ladder truck cost	-.006	-.008	.008	.007	-.025	-.036	-.728	-1.092	.468	.277	11.565	10.610
Proof of income document	.010	.007	.010	.010	.039	.027	.990	.684	.324	.495	15.967	14.587
Apartment ownership	.013	.017	.009	.008	.055	.074	1.496	2.073	.137	.040	13.621	12.256
Cleanup system	.013	.012	.008	.008	.066	.064	1.583	1.584	.116	.115	17.493	15.472

〈Table 4〉 Regression results (t-test) of C.C part II

By C.C value * **** ≥0.8 (32 variables) ***** ≥0.9 (5 variables)	Unstandardized Coefficients				Stand. Coefficients		t		p-value		-VIF	
	B		Std. Error		Beta							
	***	****	***	****	***	****	***	****	***	****	***	****
(Constant)	101.168	96.567	.981	.904			103.14	106.873	.000	.000		
CK	-.008	-	.017	-	-.019	-	-.460	-	.646	-	15.492	-
Bulgari	.044	-	.017	-	.095	-	2.597	-	.010	-	12.327	-
Credit reference	-.051	-	.013	-	-.121	-	-3.854	-	.000	-	9.185	-
Lease contract	-.001	-	.012	-	-.005	-	-.114	-	.910	-	15.298	-
Jugong apartment	.005	-	.010	-	.019	-	.497	-	.620	-	12.909	-
Latest released movie	-.026	-	.007	-	-.078	-	-3.653	-	.000	-	4.214	-
Paul Smith	-.026	-	.011	-	-.070	-	-2.355	-	.020	-	8.294	-
IBK Kiup Bank	.007	.059	.013	.015	.020	.173	.521	3.915	.603	.000	13.821	6.339
Kwangju housing prices	.052	-	.017	-	.100	-	3.003	-	.003	-	10.291	-
Minwon 24	.004	-	.017	-	.012	-	.265	-	.792	-	17.939	-
Small houses	-.013	-	.013	-	-.030	-	-1.009	-	.315	-	8.445	-
ID card authenticity check	.045	.062	.014	.017	.147	.205	3.090	3.555	.002	.000	21.110	10.790
4 bay	.027	-	.006	-	.100	-	4.340	-	.000	-	4.905	-
Galleria	-.042	-	.008	-	-.159	-	-5.132	-	.000	-	8.852	-
Nice auction	-.020	-	.008	-	-.069	-	-2.454	-	.015	-	7.314	-
Dr Apart	.023	-	.016	-	.087	-	1.426	-	.156	-	34.257	-
Restaurants	-.023	-	.008	-	-.071	-	-3.122	-	.002	-	4.796	-
Yamaha	.011	-	.010	-	.039	-	1.006	-	.316	-	13.830	-
Joinsland	-.034	-	.015	-	-.119	-	-2.232	-	.027	-	26.273	-
Brokerage commissions	.010	-	.011	-	.037	-	.859	-	.391	-	17.486	-
Used car	-.072	-.123	.011	.011	-.229	-.394	-6.496	-11.339	.000	.000	11.547	3.923
Seal certification for Sale	-.008		.006		-.043		-1.397		.164		8.606	
Pusan housing prices	-.020		.008		-.102		-2.511		.013		15.331	
Waste sticker	-.021	.022	.008	.008	-.127	.137	-2.458	2.628	.015	.009	24.595	8.864
Newlyweds home	.145	-	.066	-	.046	-	2.199	-	.029	-	4.041	-
Lotteria	.016	-	.008	-	.050	-	1.839	-	.068	-	6.876	-
Deposit return	.026	-	.009	-	.106	-	2.977	-	.003	-	11.775	-
Property views	.011	-	.004	-	.055	-	2.464	-	.015	-	4.567	-
Ladder truck cost	-.015	-	.007	-	-.069	-	-2.252	-	.026	-	8.673	-
Proof of income document	.011	-	.009	-	.044	-	1.179	-	.240	-	12.999	-
Apartment ownership	.021	.090	.007	.020	.088	.133	2.762	4.581	.006	.000	9.413	2.734
Cleanup system	.012	-	.007	-	.062	-	1.667	-	.097	-	12.625	-

between variables.[23] Therefore, a sensitivity analysis of a C.C of 0.6, 0.7, 0.8 and 0.9 has been conducted. Table 1 shows the results of the sensitivity analysis. Among the sample data of 600 search queries, 51 have shown a C.C higher than 0.6, 44 have shown a C.C higher than 0.7, 32 have shown a C.C higher than 0.8 and 5 have shown a C.C higher than 0.9.

2. Regression Analysis

This chapter deals with a detail explanation of Step. 4 of Figure 1. Step 4.1 deals with the sensitivity analysis presented in the previous chapter where the search query data has been divided into groups based on the C.C values. As shown in Step 4.2, regression analysis for each of the selected groups of C.C has been applied. The HSI has been set as the dependent variable and search queries have been as the independent variables. ANOVA (Analysis of variance) has also been applied to each model. The results are shown in Table 2. As shown in Table 2, all the p-values of the models are lower than 0.05. That indicates that all the models are statistically verified for use in

23) Devore, J., *Probability and Statistics for Engineering and Science*, Richard Stratton, 8th ed., 2012. pp.1-712.

〈Table 5〉 Model summary of each group divided by C.C

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
C.C ≥ 0.6 (51 variables)	0.992	0.984	0.979	0.6701
C.C ≥ 0.7 (44 variables)	0.991	0.983	0.978	0.6873
C.C ≥ 0.8 (32 variables)	0.991	0.981	0.978	0.6968
C.C ≥ 0.9 (5 variables)	0.968	0.938	0.936	1.1767

a significance level of α=0.05. The t-tests for each of the coefficients of the regression model are shown in Tables 3, 4.

As shown in Step. 4.3, The R square and VIF values have been calculated. The VIF values are shown in Tables 3, 4 and the R square values are shown in Table 5. Table 5 shows the model summary of each model showing the R square and adjusted R square values. The R square and adjusted R square values of all the models is higher than 0.9. That indicates that the models can explain more than 90% of the HSI data.

As shown in Table 5, the R square value of the model where C.C≥0.6 have shown the highest value of 0.992. The models of C.C≥0.7 and C.C≥0.8 have shown a similar value of 0.991. However, as shown in Table 3, numerous independent variables have shown a VIF value higher than 10 which indicates that a multicollinearity exists in those models. Moreover, numerous independent variables have shown not significant for analysis because the p-value of the t-test have shown values higher than 0.05. A noticeable feature of the regression analysis is that the R square value increases when more independent variables are inserted. Thus, in this study,

〈Table 6〉 Model summary of the final model

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
Final	.966	.934	.933	1.2100

〈Table 7〉 ANOVA table of the final model

Model		Sum of Squares	df (degree of freedom)	Mean Square	F	p-value
Final	Regression	4204.082	4	1051.020	717.847	.000
	Residual	297.218	203	1.464		
	Total	4501.300	207			

〈Table 8〉 Regression results (t-test) of the final model

	Unstandardized Coefficients		Standardized Coefficients	t	p-value	VIF
	B	Std. Error	Beta			
(Constant)	97.379	.899		108.329	.000	
IBK Kiup Bank	.074	.015	.217	4.990	.000	5.830
Used car	-.133	.011	-.425	-12.321	.000	3.667
Waste sticker	.040	.007	.246	5.658	.000	5.816
Apartment ownership	.106	.020	.155	5.340	.000	2.604

it is recommended to use only the model of C.C≥0.9 which only one variable has shown multicollinearity. As shown in Step .4.4, regression analysis has been applied again to the group of C.C≥0.9 using the forward selection method to exclude the variables "ID card authenticity check" that has a VIF value of 10.790. The results of the analysis are shown in Tables 6, 7, 8.

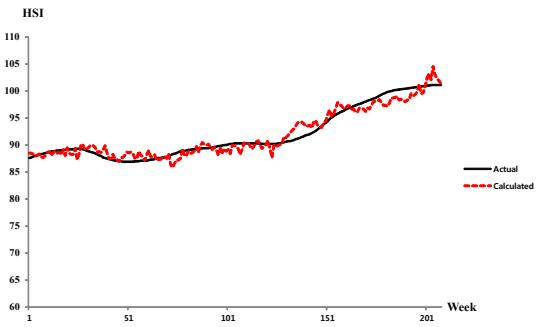
The final model has been derived as shown on the Tables 6, 7, 8. High values of R square and adjusted R square of 0.933 and 0.932 have been obtained. The ANOVA test has also shown statistical significance based on the p-value (0.000) which is smaller than 0.05 (α=0.05). Based on Table 8, the final model can be expressed as equation (2). Figure 5 shows the comparison graph of actual HSI values with the values calculated based on the final model.

V. Discussion and Applications

As shown in Figure 6, the model successfully predicts HSI using the independent variables "IBK Kiup Bank", "Used Car", "Waste Sticker" and "Apartment Ownership".

$$HSI = 97.379 + 0.074 \times (IBK\,Kiup\,Bank) - 0.133 \times (Used\,Car) + 0.040 \times (Waste\,sticker) + 0.016 \times (Apartment\,ownership) \quad (2)$$

〈Figure 6〉 Comparison graph of actual and calculated HSI



The final selected model as shown in equation (2) can be used for predicting HSI within the range of 86.9 to 101.1 which are the minimum and maximum values of HSI. However, the value of HSI has been investigated to have a potential to grow much higher than 101.1. In such a case the methodology should be applied once more to a newly collected data set. Therefore, in future studies, a methodology that can overcome such limitation must be developed. MAE (Mean Absolute Error) and NMSE (Normalized Mean Squared Error) has been calculated based on equation (3) and (4).[24)25)]

$$NMSE = \frac{1}{n}\sum_{t=1}^{n}\frac{(Actual_t - \mathrm{Predicted}_t)^2}{\sigma^2\,of\,predicted} \quad (3)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|Actual_t - \mathrm{Predicted}_t| \quad (4)$$

24) Anwar, S., and Mikami, Y., "Comparing Accuracy Performance of ANN, MLR, and GARCH Model in Predicting Time Deposit Return of Islamic Bank. International Journal of Trade," *Economics and Finance*, 2011, vol.2, no.1, pp.44.

25) Chun, H, "A Study on the Volatility and Spillover Effect of Housing Sales, Chonsei, and Monthly Rent Market Using GARCH, EGARCH Model", *Korea Real Estate Academy Review*, 2015, vol.62, pp.218-232.

where,

$\sigma^2 of predicted$ is the total variance of the predicted values

It has been calculated that the MAE and the NMSE of the final model are 0.967 and 0.070 respectively. It has been analysed in the previous chapter that the R square and adjusted R square values are higher than 0.9. Accordingly, the MAE and NMSE values have also shown low values as expected. However, it must be stated that a data such as the HSI that has a low standard deviation (4.666), generally, results in a very well fitted prediction models. Therefore, despite the excellent numerical results of this study, a new method that is capable of suggesting lower MAE and NMSE values must be suggested in future studies. The final model suggested in equation (2) uses only the variables "IBK Kiup Bank", "Used Car", "Waste Sticker" and "Apartment Ownership" due to the existence of multi-collinearity between variables. Also, it has been shown that as more independent variables are inserted into the regression analysis the R square value increases. However, using numerous independent variables (51 variables of C.C≥0.6 for example) is expected to be more time consuming and complicated for the actual users of the suggested methodology. Therefore, methodologies of appropriate grouping of variables and extracting simple representative values must be suggested in future studies that are capable of consistent data updating and suggesting simple regression models that would be easily used in the field.

Ⅵ. Conclusion

This study analyses the relationship between HSI and search queries for suggesting a reliable prediction model. The prediction model was suggested based on a multiple regression analysis method. R square and adjusted R square values have been calculated as 0.934 and 0.933 respectively. However, it has been investigated that despite the high values of R square and adjusted R square, the prediction results must be improved based on the calculation of MAE and NMSE. Analysis conducted in such data sets must be checked not only with R square but also with MAE to compare with the variance of the actual data. The suggested prediction model, however, can be used as a reliable reference to understand the pattern of HSI changes that is expected to be used in establishing efficient economic strategies in the planning and preliminary design phase of construction projects. This study focuses on suggesting a reliable HSI prediction model, however, it is expected that the proposed methodology can be used in various research fields by using search query data. The most distinguishable aspect of this study is the suggestion of a new approach in which a big data analysis method expanded from conventional cause-effect relationship. It is expected that other various and reliable prediction models will be

applicable to the construction industry, which is considered one of the most conservative industries owing to its limitations in the practical application of decision-making tools.

References

Ahn, J.J., Byun, H.W., Oh, K.J., and Kim, T.Y. "Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting", *Expert Systems with Applications,* 2012, vol.39, no.9.

Anwar, S., and Mikami, Y., "Comparing Accuracy Performance of ANN, MLR, and GARCH Model in Predicting Time Deposit Return of Islamic Bank. *International Journal of Trade, Economics and Finance,* 2011, vol.2, no.1.

Bang, K., *Real Estate Terms Dictionary,* Buyeonsa. 2011.

Beracha, E., and Wintoki, M.B., "Forecasting residential real estate price changes from online search activity", *Journal of Real Estate Research,* 2013, vol.35, no.3.

Bollen, J., Mao, H., and Zeng, X., "Twitter mood predicts the stock market," *Journal of Computational Science,* 2011, vol.2, no.1.

Choi, H., and Varian, H., "Predicting the present with Google Trends", *Technical report, Google.* 2009

Choi, H., and Varian, H., "Predicting the present with Google Trends," *Economic Record,* 2012, vol.88, issue s1.

Chun, H, "A Study on the Volatility and Spillover Effect of Housing Sales, Chonsei, and Monthly Rent Market Using GARCH, EGARCH Model", *Korea Real Estate Academy Review,* 2015, vol.62.

Da, Z., Engelberg, J., and Gao, P., "In Search of Attention," *Journal of Finance,* 2011, vol.66, no.5.

Devore, J., *Probability and Statistics for Engineering and Science,* Richard Stratton, 8th ed., 2012.

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,* 2009, vol.457, no.7232.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J., "Predicting consumer behavior with Web search," *Proceedings of the National academy of sciences,* 2010, vol.107, no.41.

Joseph, K., Wintoki, M.B., and Zhang, Z., "Forecasting Abnormal Stock Returns and Trading Volume Using Investor Sentiment: Evidence from Online Search," *International Journal of Forecasting,* 2011, vol.27.

Kim, D. and Yu, J., "A Dynamic Relationship Between Internet Search Activity, Housing Price, and Trading Volume", *Korea Real Estate Review,* 2014, vol.24, no.2.

Kim, H., and Lee, S., "Analysis on the Consumer Behavior and Location Characteristics based on Big Data for Department Store-Type Discount Store Development", *Korea Real Estate Academy Review,* 2015, vol.63.

Ko, Y., Han, S., and Lee, S., "Preliminary analytic methodologies for home sales index assessment based on big-data analysis", *Review of Real Estate and Urban Studies,* forthcoming 2015, vol.8, no.2.

Lee, J., and Lee, C., "Analysis on the Determinants of Korea Housing Price Index in Unstable Housing Market by Volatility of the Housing Price", *Korea Real Estate Academy Review,* 2014, vol.59.

Naver Trend, Search Queries available from Internet: http://trend.naver.com. 2016.

Repkine, A., and Song, S., "A Spatial Approach to the Hedonic Pricing of Apartment Attributes", *Korea Real Estate Academy Review,* 2015, vol.63.

Sim, S., "Panel Analysis of Relationship between House Sales Prices and Trading Volume", *Korea Real Estate Academy Review,* 2015, vol.63.

Tuna, C., "New Ways to Read Economy," *The Wall Street Journal,* 2010, April 8. (http://www.wsj.com/articles/SB10001424052702303395904575158030776948628)