Predicting sales by online searching data keywords based on text mining: Evidence from the Chinese automobile market

Yi Li¹, Liangru Yu and Rui Wen

School of Economics and Management, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
E-mail: ¹ liyi@cqupt.edu.cn

Abstract. Online searching data reflects consumers’ real footprints in information collection and purchase decision-making processes, which is greatly valued in understanding their needs. This paper which is at the background of China's automobile market, studies the relationship between online searching data and automobile sales through approaches that differ from existing research to extract keywords. First the online searching data keywords are determined, primarily by using text-mining technology to extract them, and specifically: i) Jieba was used to tokenize crawled automotive forum posts’ text into segmented words; ii) All word-segmented Chinese corpus were segmented into word vector space by Word2vec model; and iii) Similar keywords were discovered by calculating the word vector’s similarity indexes. A fixed effect model was then built based on 108 months of long panel data. Finally, comibng with panel vector autoregressive model (PVAR), we used rolling window to predict Chinese automobile sales from January to December 2015. The empirical results demonstrate that: a long equilibrium exists between online searching data and automobile sales; our regression model can explain 76% of the variance. The holdout analysis suggests that online searching data can be of substantial use in forecasting Chinese automobile sales.

1. Introduction

Due to the emergence of such big data sources as Google Trends, monitoring market participants’ behavioral tendencies and patterns will become increasingly cost-effective. Ginsberg et al. first found a strong connection between Google Trends’ search volume data of flu-related terms and the incidence of flu in the United States[1]. The value of Google Trends had been previously verified in many areas, such as forecasting the new product[2], oil[3]. However, as Baidu is the most popular search engine in China[4], using the Baidu Index to study Chinese socioeconomic behavior more closely parallels the actual situation.

Jiang et al. used the Baidu Index to fit and forecast Shanghai City’s new residential price index[5]. Ren and Cui empirically tested the relationship between Beijing’s search query volume and monthly tourist volume by calculating the correlation coefficient, determining the leading order, and stepwise synthesizing keywords data, and Baidu’s synthetic search query volume[6]. Yuan et al. established a theoretical, basic framework from the perspectives of online search keywords and automobile sales, and a comprehensive weighted method was used to extracted keywords to predict automobile sales[7].

However, regarding existing research in China, the use of online searching data to predict consumers’ demands is still a new research area without a systematic research system, this deficiency is further noted, as follows. (1) Controversy still exists as to how to select and synthesize online search keywords. Most research involves directly specified keywords, or is combined with Baidu’s automatically
generated keywords to extract an index, without considering if the keywords can represent the users’ online search behavior. (2) Prediction models adopt a time-series static regression or a short-panel data model, which is not conducive to controlling individual differences. This is also not conducive to accurately reflecting and depicting the dynamic changes in the influence of online searching data on sales with the passage of time. This paper has improved upon the following aspects to compensate for this deficiency in existing research:

(1) Keywords extraction: Based on text-mining technology, systematically extract keywords that accurately reflect consumers’ online search behavior to reflect their accurate demand. First, the cosine distance between word vectors is calculated to discover similar predictive queries by using Python to crawl automotive forum posts, Jieba to tokenize the forum post text corpus into segmented words, and Word2vec to train the word-segmented Chinese corpus into a vector form. Second, similar keyword frequency statistics are obtained from forum posts’ text. Third, the word with maximum correlation coefficient between the variables at different lags is selected as the keyword.

(2) Model specification: Build a fixed-effect model using long panel data, based on Chinese automobile sales data and Baidu Index data from 2007 to accurately reflect the dynamic changes in online search data’s influence on sales with the passage of time. Further, a rolling window is used to predict car sales in 2015.

This paper chooses automobile as research object since the automobile has complex attributes and needs large capital investment. Consumers will investigate and assess all attributes carefully in the purchase of automobile. Also automobile is one of the most commonly used research object in the research of consumer’s external search[8].

This paper’s primary work involves presenting a structured process to determine the search keywords in online searching data based on the latest research achievement on word vectors in natural language processing. This is then applied to Chinese automobile sales research, which provides a reference to research online searching data in other areas.

2. Methodology

2.1. Data resources

2.1.1. Automobile sales. The automobile sales data in this paper comes from Sohu’s website: http://db.auto.sohu.com/cxdata/, which is updated monthly. A total of 108 months, from January 2007 to December 2015, were selected to study the long-term relationship between online searching data and automobile sales. The sustainable availability of sales data is considered, in that automobile models without sales for consecutive 12 months are excluded. Finally, 55 models’ monthly sales data is selected from the Chinese market.

2.1.2. Online searching. This paper’s online searching data comes from Baidu Index data. Baidu is the world’s largest Chinese search engine. Baidu's market share in China reached 82.3% by the third quarter of 2015, far more than the subsequent “Google China” (7.9%), Sogou (4.8%), or 360 Search (3.8%)[4]. Meanwhile, Baidu Index data is a free analysis service that amasses data based on Baidu’s web searches, which can reflect the “user attention” of different keywords during a period. “User attention” is based on Baidu’s web searching volume, and takes keywords as statistical objects, representing each keyword’s search frequency in Baidu’s web searches, and is updated once daily.

2.2. Keywords selection

Several different ways of keyword selection were adopted in empirical researches. Seebach et al. used the names of auto manufacturers, Volkswagen, general motors etc., to collect data including all the brands and models of the manufacturers[9]. However, to select keyword as this way may not truly reflect consumers’ purchase intentions. Because the search requests containing brands or manufacturer’s name may include the non-transactional information needs. For example, some search requests only aimed at
company’s management or recruitment information which was clearly not related to purchase. According to a survey, when consumers considered buying a car, the name of the car was the most commonly used as a keyword of online searching, followed by price, brand and manufacturer's name[10]. The consumers searching certain specific models have clearer purchase goal and stronger purchase intentions than the consumers taking brands or manufacturers’ names as searching keywords. Therefore, this paper proposes model name or compound words like brand with model name to be basic search terms.

However, consumers may consider using some synonyms for the model name as keywords to search for information due to the non-standard use of language and impact of individual subjectivity. For example, consumers may use “Small4L” to replace the “Audi A4L.” Synonyms’ semantic functions and rules are used to comply with the following key points: Substitution between synonyms always occurs in a similar context, which means that the appearance of different synonym items in a group of synonyms often implies the similarity of context. Therefore, this can be regarded as a sign of the same “topic context”[11]. A mathematical structure is always used in text mining, such as “vector” to represent a word. Each of the vector’s dimensions represents a semantic or syntactic feature of the word to express its meaning within its context as much as possible. Google Corporation proposed a training word-vector model in 2013, with the word “vector,” as the characteristics of the input word can effectively and quickly find the words that approximate the specific word. Word2vec is based on deep learning, input for document collection, and output, as the word “vector,” through calculating the distance between the word “vector” (where the distance can be cosine similarity or Euclidean distance), can discover the neighbor of a given word[12]. As Word2vec can characterize each word to a k-dimensional real-valued vector, we will obtain similar words by calculating the similarity in this vector. We have an innovative application to find the synonyms of the basic search terms’ synonyms, and to supplement this research’s achievements in deep learning and word vectors in natural language processing using Word2vec. This considers that Word2vec requires a large number of text training corpus as input. Therefore, we chose Chinese as the training text corpus, and the largest automobile forum with the highest Chinese user activity: Autohomes[13]. The process to obtain the synonyms is found in Step 2.

In the next step, we will place possible keywords into the Baidu Index; to retrieve the search index, we select the keyword with highest search index, similar to the work of Hu et al.[14]. While for cannot be determined for search index rankings, we calculate the person correlation coefficient under different lags in Step 4 for a final determination. The following notes all processes in detail to determine this paper’s online searching keywords.

(1) Determine basic search terms. This paper takes model names and compound words, such as brands with model names, from the Sohu automobile channel (http://db.auto.sohu.com/cxdata/) as the first basic search terms (e.g., “A4L” and “AudiA4L”).

(2) Similar words are obtained from basic search terms. 1) Independent development of Python allows for a crawl of forum posts’ content from web pages with the modeled forum topics in the Autohomes forum. 2) Jieba is used to tokenize the forum posts’ text corpus, and then convert it into segmented words as a post-text feature. 3) Word2vec is used to train the word-segmented Chinese corpus, and transform this into word vectors. 4) Based on the cosine similarity algorithm and distance algorithm, Word2vec distance is used to calculate the cosine distance (or cosine similarity) between the basic search term vector and other term vectors, with 0.5 set as the threshold to return terms similar to the basic search term, or the synonymy terms, after sorting (this does not consider the extension of the term, such as “A4L fuel consumption”). The synonymy terms are then iteratively trained for the second round. This process repeated for several rounds, and 452 total terms were obtained. Duplicate terms were removed using a structured query language (SQL), ultimately resulting in 318 search terms.

(3) Select forum high-frequency terms. Each term’s frequency is estimated in the forum’s text, and terms with high frequency are selected as target terms in the Baidu Index’s data. The brand name is added to the target search keywords for still ambiguous terms. For example, for “JinGang,” the target search keyword will be “JiLiJinGang.” Similar examples include “YuYan,” “BeiDouXing,” and “GaoErFu,” among others.
(4) Determine final search terms. Target terms are searched in Baidu Index data, and the highest-ranking words in the Baidu Index are chosen as keywords. If the words cannot be determined from the ranking, words are chosen with the highest Person correlations during different lags, from 0 to 6, as searching keywords. Finally, 55 models are determined as unique searching keywords from the Baidu Index data.

Pearson correlations for each final keyword are then calculated between the word and sales, during lags from 0 to 12. Table 1 displays the lag periods for keywords based on the maximum Pearson correlations between their Baidu Index data and sales. Lags are generally concentrated in periods of 0 to 2, and a majority of the lag period is 0. Although consumers may in reality have several months to perform an information search and evaluation before their final purchase, the searching number and strength are relatively small, and until a month before purchase, their search quantity and strength will reach a critical point.

Table 1. Keywords and their lags

| Key word | Lag | Key word | Lag | Key word | Lag |
|----------|-----|----------|-----|----------|-----|
| Yuan Jing 0 | POLO 0 | Feng Tian Huang Guan 1 | Ji Li Zi You Jian 0 | 0 |
| Jun Yue 0 | Kai Yue 1 | Bao Ma 3 Xi 0 | Ling Mu Yu Yan 2 | 2 |
| Tu Sheng 0 | Jing Cheng 0 | Ma Zi Da 6 | Da Zhong Gao Er Fu 0 | 0 |
| Xuan Yi 0 | Ao Tuo 0 | Feng Tian Wei Chi 0 | Ling Mu Bei Dou Xin 0 | 0 |
| Si Yu 0 | Jie Da 0 | Ben Tian CRV 0 | Ben Teng B70 2 | 2 |
| Rui Zhi 2 | Tian Lai 4 | San Ling Ge Lan 0 | JIn Bei Ge Rui Si 7 | 7 |
| Tu An 0 | Ao Dei Sai 0 | Sai La Tu 0 | Zhong Hua Jun Jie 0 | 0 |
| GL8 0 | Fu Mei Lai 0 | Bao Ma 5 Xi 0 | Qi Rui QQ3 0 | 0 |
| Su Teng 0 | Hua Guan EX 0 | Sai Ou San Xiang 2 | Fu Ke Si San Xiang 4 | 4 |
| Ya Ge 0 | Te La Ka 0 | Jiang Hua Rui Feng 0 | Di Ao A4L 0 | 0 |
| Fei Du 0 | Sheng Da Fei 0 | Pu La Duo 0 | Wei Zhi San Xiang 0 | 0 |
| Le Chi 2 | Pu Li Ma 1 | Ji Li Jing Gang 3 | Li Ya Na Liang Xiang 0 | 0 |
| Jun Wei 0 | Kai Mei Rui 0 | San Ling Lan Se 0 | Di Ao A6L 0 | 0 |
| Qi Da 0 | Pa La Ding 0 | Bi Ya Da F3 0 | Ling Mu Yu Yan 2 | 2 |

Note: We used Pinyin to interpret Chinese characters.

3. Relationship between Baidu Index data and sales

3.1. Unit root tests

Assume sales (or “S”) as the dependent variable and Baidu Index data (or “B”) as the independent variable. The variables are converted to logarithmic form (InS, InB) in this paper to reduce heteroscedasticity’s impact on the test, and to consider the skewed distribution of sales and Index data. One advantage in using a logarithmic form is that the prediction of online searches can be explained by a change in percentage, rather than absolute value.

As all this paper’s samples involve panel data, unit root and co-integration tests are required before building an econometrics model, as well as a time series to ensure variables’ stationarity and avoid spurious regression. This paper uses LLC, PP, IPS and ADF to perform unit root tests. The result in Table 2 demonstrates that sales and online searching data are stable in all cases. As a co-integration test investigates if co-integration exists among a linear combination of non-stationary series[15], this is not performed in this paper for real sales and the online search index, but rather, directly into Granger’s causality test.

3.2. Granger causality test

Real-life consumers always search for information before buying a car, so online searching occurred before their purchase, which means the Baidu Index data causes sales. It is necessary to conduct a Granger causality test to verify whether the actual data can support this conclusion, and a precondition of the test is the stationarity of data. The lags are 0 to 2 months from online search to purchase. As this paper focuses on online searching data as a predictor of sales, the same period’s influence is not
considered. The results of the Granger causality test, in which lags were set to 1 to 2, are displayed in Table 3.

According to the result, a bidirectional and interactional relationship exists between online searching data and sales in the case of 1 and 2 lags. However, when sales are predicted by the Granger causality test, the initial information from online searching data reduces more mean square errors (MSEs) than in the other case.

| Table 2. Unit root tests of panel data |
|--------------------------------------|
| Variables | \( \ln B \) | \( \ln S \) |
| LLC       | (A)     | -5.675*** | -9.225*** |
|           | (A,T)   | -7.632*** | -21.590*** |
| PP        | (A)     | 214.074***| 757.240*** |
|           | (A,T)   | 368.976***| 1075.280***|
| IPS       | (A)     | -1.596*  | -13.977*** |
|           | (A,T)   | -8.176*** | -23.765*** |
| ADF       | (A)     | 202.397***| 563.217*** |
|           | (A,T)   | 346.924***| 903.453*** |

Note: ***, **, and * represent significance at the 1%, 5%, and 10% levels, respectively; (A) represents the including intercept, and (T) is the trend in unit root tests.

| Table 3. Granger causality test |
|---------------------------------|
| Hypothesis | F test | Probability | Result | F test | Probability | Result |
|------------|--------|-------------|--------|--------|-------------|--------|
| Lag 1 month | Lag 2 month |
| \( \ln B \) is not the cause of \( \ln S \) | 44.101 | 3E-11 | reject*** | 18.245 | 1.E-08 | reject*** |
| \( \ln S \) is not the cause of \( \ln B \) | 11.776 | 0.000 | reject*** | 6.430 | 0.002 | reject*** |

Note: ***, **, and * represent significance at the 1%, 5%, and 10% levels, respectively.

3.3. Establishment of panel data model

Panel data is used to analyze the relationship between online search and sales, and an F-test is required to analyze which model is suitable: a mixed estimation or fixed effects model. This paper uses a fixed effects model according to the results of this test. The primary reason for endogeneity in the regression model involves omitting important relevant variables. Once the omitted variable is related to both the dependent and objective independent variables, the independence hypothesis of the objective independent and random perturbed variables is no longer valid. At this time, if the least-squares estimation is used, the objective independent variables’ estimation is biased. This paper considers this by using a fixed-effects model to estimate panel data to control for endogeneity. One significant advantage of panel data involves using the fixed-effects model to control cross-sectional variation. Further, the Hausman Test’s result rejects random-effects models. As lags exist between dates and online searching, this paper takes the previous one and two online searches as the independent variables to test if the previous one and two month online searches can predict sales. The model for model \( i \) and time \( t \) is:

\[
\ln S_{brd,month} = \beta \ln B_{brd,month} + \mu_{brd} + \varepsilon_{brd,month} \quad (brd = 1, \ldots, 55; month = Jan2007, \ldots, Dec2015)
\]

\[
\ln S_{brd,month} = \log S
\]

\[
\ln B_{brd,month} = (\log B1, \log B2)
\]
Cross-sectional weighting and the generalized least squares method (GLS) are adopted to estimate the model and reduce the heteroscedasticity of the cross-sectional data, and the results are noted in Table 4. This table indicates that the Baidu Index’s data can predict sales, and the model can explain the 76% variance in sales.

Table 4. The regression results of panel data

| Variable | Meaning                                | Hausman test | Coefficient | Constant | R²   |
|----------|----------------------------------------|--------------|-------------|----------|------|
| logB1    | Online searches ahead 1 month           | 29.702***    | 0.207***    | 6.638*** | 0.762|
|          |                                        | (26.195)     | (118.296)   |          |      |
| logB2    | Online searches ahead 2 months          | 29.405***    | 0.206***    | 6.636*** | 0.761|
|          |                                        | (25.858)     | (116.652)   |          |      |

Note: ***, **, and * represent significance at the 1%, 5%, and 10% levels, respectively; the values in parentheses note the t-statistics of the parameter estimations.

4. Panel vector autoregressive model (PVAR) estimation

4.1. Establishment of vector autoregressive model

Since the panel data is stable, it can be considered to establish a vector autoregressive model for panel data (Panel VAR, PVAR). In order to determine the appropriate lag length p of the model, 12 is selected as the maximum possible lag order. The order of the model is determined according to the criterion that the AIC, SC, and HQ information values are the smallest[16]. The results show that the optimal lag order of AIC is 12, but the number of HQ and SC are 5. According to majority rule[16], the lag order of the VAR model can be defined as 5th order. Then establish the VAR model, and the estimated results are shown in equation (2), whose goodness of fit is $R^2_{lnS} = 0.821$, $R^2_{lnB} = 0.981$. All the characteristic root of the VAR model are located in the unit circle, the structure of model is stable, and the fitting effect is good.

$$
\begin{align*}
\ln S_i & = 0.197 + 0.623 \ln S_{i-1} + 0.149 \ln S_{i-2} + 0.160 \ln B_{i-1} - 0.035 \ln B_{i-2} \\
\ln B_i & = 0.001 + 0.847 \ln B_{i-1} + 1.820 \ln B_{i-2} - 0.045 \ln B_{i-3} + 0.076 \ln S_{i-3} - 0.003 \ln B_{i-3} + 0.004 \ln B_{i-4} + 0.075 \ln B_{i-4} - 0.04 \ln B_{i-5} + 0.062 \ln B_{i-5} \\
\end{align*}
$$

4.2. Impulse response function and variance decomposition

Since the vector autoregressive model (VAR) is a non-theoretical model, its coefficients are difficult to explain. When analyzing vector autoregressive model (VAR), impulse response function and variance decomposition are generally used for further analysis. The impulse response functions of automobile sales and online search data are shown in Figure 1. After applying a shock of the standard deviation to the automobile sales volume $\ln S$ and the online search data $\ln B$, the whole process of complex relationship and mutual influence of two variables is described.

As can be seen from the impulse response function of the automobile sales in Figure 1 (a), the positive shock of a standard deviation from the online search data has a significant effect on the increase in automobile sales, and this effect has a long lasting effect. Specifically, due to an accidental increase in online search data, automobile sales in the first three months have increased significantly and reached the highest level, and slowly declined from the fourth month to the sixth month. After that, it maintained a slow growth trend and basically stabilized. It can be seen from Figure 1 (b) that after a positive shock on automobile sales in the current period, the online search data reached its peak in the fifth period after a small fluctuation in the first four periods, and then began to grow steadily, but its overall impact effect is much smaller than the codirectional impact of Internet search data on automobile sales.
Note: The solid line indicates the impulse response of a standard deviation innovation, and the dashed line indicates the addition or subtraction of the standard deviation of the two sides of the response pulse function image.

Figure 1. Impulse response function diagram of automobile sales and online search data

The impulse response function describes the impact of the shock of an endogenous variable in the VAR model on other endogenous variables. The variance decomposition is to further evaluate the importance of different structural shocks by analyzing the contribution of each structural shock to changes in endogenous variables (usually measured by variance)[17]. As with time series data, the variance decomposition of the panel data gives relatively important information for each random disturbance term that affects the variables in the VAR model. The variance decomposition of each main variable is shown in Figure 2 and Figure 3.

According to the results of variance decomposition, regardless of the contribution rate of automobile sales, the contribution rate of online search data to automobile sales has gradually increased from the first period, and to the maximum in 12th period, among which the first four periods show significant growth, then maintained a slow growth trend, and basically stabilized after the eighth period. However, the variance decomposition of the network search data is mainly composed of itself, and automobile sales have a slight impact on it, which is almost negligible. This is consistent with the results of the impulse response function and the Granger causality test, that is, online search data can drive automobile sales, and automobile sales have little impact on network search data.

4.3. Out-of-sample prediction of vector autoregressive model (VAR)
In-sample analysis has proven that online search data has a strong ability to interpret automobile sales. Next we will examine the predictive power of online search data by out-of-sample analysis of vector autoregressive models (VAR). The test analysis method is similar to that of Seebuch et al.[9], that is, the
window of the rolling window is used to move backward to predict the sales volume from January to December 2015. Specifically, the model’s parameters are estimated by using the first 96 months’ observational data to predict sales in the 97th month. The observational data from months 2 to 97 were then used to re-estimate the model’s parameters and predict sales in the 98th month, and so forth.

Figure 4 and Figure 5 are the comparison charts of the actual sales data and predicted sales data of the models Terracan and Santa Fe, in which lnS is the logarithm of the actual sales data, and lnSF is the predicted value. It can be seen from the comparison chart that the online search data is well combined with the automobile sales data, and the model prediction effect is very good.

![Figure 4. Comparison of actual sales data and forecast data of Santa Fe](image)

![Figure 5. Comparison of actual sales data and forecast data of Terracan](image)

According to the method proposed by Hydmen and Koehler[18], the article chooses the mean absolute error (MAE) as an indicator to evaluate the predictive power. The unit of the mean absolute error is consistent with the unit of the dependent variable and is easy to interpret. In addition, the article calculates the root mean square error (RMSE) of the predicted value, because this index is more sensitive to outliers than the average absolute error. Using them together can bring a more comprehensive evaluation of the model's predictive power. The calculation methods of mean absolute error and root mean square error are as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \bar{Y}_i| \quad RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2}$$

The results show that MAE=0.431 RMSE=0.756, the model prediction effect is very good. The model’s goodness of fit and forecast accuracy are improved by adding Baidu Index data. The result once again reflects the prediction effectiveness of the online search data.

5. Conclusion

This paper first proposes a set of structured processes to determine the keywords of the online search data, and then make an empirical study on the relationship between the online search data and China's automobile sales with a fixed effect model of the panel data by using unit root test and the Granger causality test. Finally, based on the panel vector autoregressive model (PVAR), the impulse response function and variance decomposition analysis are carried out. The prediction effectiveness of the online search data is tested by rolling window. Concluded as follow:

(1) Granger causality test shows that in the case of 1 and 2 periods ahead, there is a two-way interaction causal relationship between online search data and automobile sales, but the advance of online search data contributed more to the reduction of mean square error.

(2) The panel data model shows that by establishing a regression model of fixed effects, the online search data can explain 76% of the variance of sales, that is, there is a long-term equilibrium relationship between online search data and automobile sales.

(3) The impulse response function shows that: online search data plays a significant role in promoting automobile sales, but automobile sales have little impact on online search data.
(4) The out-of-sample prediction results of the vector autoregressive model (VAR) show that the mean absolute error and root mean square error of the predicted values are 0.431 and 0.756, respectively, indicating that the online search data has a strong predictive effect on automobile sales.

Due to the research of the article, we can predict the changes of automobile sales in the Chinese market without excessive reliance on historical data, and provide a reference to automobile companies and government departments.

**Acknowledgments**

This work was supported by the Undergraduate Scientific Research Training Program of Chongqing University of Posts and Telecommunications.

**References**

[1] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L. (2009) Detecting influenza epidemics using search engine query data. Nature, 457: 1012–1014.

[2] Chumnunpan, P., Shi, X. (2019) Understanding new products’ market performance using Google Trends. Australasian Marketing Journal. In Press.

[3] Yu, L., Zhao, Y., Tang, L., Yang, Z. (2019) Online big data-driven oil consumption forecasting with Google trends. International Journal of Forecasting. 35: 213–222.

[4] Iresearch. (2016) 2015Q3 online advertising overall and search engine market development. http://news.iresearch.cn/zt/256851.shtml#a2.

[5] Jiang, W. j., Lai, Y. F., Wang, K. (2016) The correlation research of property prices based on Baidu Index. Statistics and Decision, 2: 90-93.

[6] Ren, L., Cui, L. J. (2014) Predicting domestic tourism traffic based on the online search data: as example of Beijing tourism traffic. Inquiry into Economic Issues, 4: 67-73.

[7] Yuan, Q. Y., Peng, G., Liu, Y., Lv, B. F. (2011) Predicting automobile sales based on online search keywords data. Management: the academic version, 1: 12-24.

[8] Klein, L. R. (1998) Evaluating the Potential of Interactive Media through a New Lens: Search versus Experience Goods. Journal of Business Research, 41: 195-203.

[9] Seebach, C., Pahlke, I., Beck, R. (2011) Tracking the digital footprints of customers: how firms can improve their sensing abilities to achieve business agility. In: European Conference on Information Systems, Helsinki, 258.

[10] China Internet Network Information Center (CNNIC). (2013) 2012 Chinese netizens consumer behavior survey report. car. http://www.cnnic.cn/hlwfzyj/hlwzbg/dzswbg/201301/t20130116_38523.htm.

[11] Chaves-González, J. M., Martínez-Gil, J. (2015) Evolutionary algorithm based on different semantic similarity functions for synonym recognition in the biomedical domain. Knowledge-Based Systems, 37: 62-69.

[12] Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013) Efficient estimation of word representations in vector space. Computer Science.

[13] 199IT. (2016) Similarweb: China’s mainstream car website ranking report in January 2016. http://www.199it.com/archives/439042.html.

[14] Hu, Y., Du, R. Y., Damangir, S. (2014). Decomposing the Impact of Advertising: Augmenting Sales with Online Search Data. Journal of Marketing Research, 51: 300–319.

[15] Gao, T. M. (2009) Econometric analysis and modeling: eviews applications and examples (The second edition). Tsinghua University Publishing, Beijing.

[16] Yi, D. H. (2014) Data analysis and EViews application. China Renmin University Press, Beijing.

[17] Gao, T. M. (2009) Econometric analysis methods and modeling: EViews applications and examples (2nd edition). Tsinghua University Press, Beijing.

[18] Hyndman R J, Koehler A B. (2005) Another look at measures of forecast accuracy. International Journal of Forecasting, 22: 679-688.