Monitoring Research on International Chinese Education Market Based on Web Searching Data

Chong ZHANG¹, Xue-Yu BAI¹ and Xian-Yu LI¹

¹College of Information Science, Beijing Language and Culture University, Beijing, China
zhangchong@blcu.edu.cn

Keywords: Internet Web Search Data, Google Trends, Composite Index, Panel Data.

Abstract. This paper uses the number of foreign students studying in China as a measure of Chinese learning heat. On the basis of building theoretical frame of web search data and the number of Chinese learners, this paper selects initial search engine keywords from interviews of 25 foreign students from 15 countries, and do Keyword expansion work by using Google and Baidu, composes the keywords into composite index which can be used in the regression model of this paper, processing panel data and modeling by using Eviews software, finally analyzes the relationship between the numbers of foreign students studying in China and the corresponding composite index. The fitting degree of model is 94.5%.

Introduction

International Chinese education is an important part of China's education market. It is of great significance to capture the fluctuation of demand in the Chinese education market in different countries and regions in terms of education resource allocation, teacher training, and Chinese communication strategy formulation. At present, the quantitative research on the fluctuation of international Chinese education market demand adopts the sampling survey statistical method. This method has the limitations of timeliness (the lag of statistical survey), subjective (susceptible to human factors), and small sample size. The development of Internet big data and information technology has provided new research ideas and data foundation for the quantitative measurement and monitoring of international Chinese education market demand.

The Internet has become the main channel for people to obtain information. When people from all over the world are interested in learning Chinese, they will browse and search for relevant information on the Internet, such as Chinese learning videos, HSK test materials, Chinese universities, applications for studying abroad, Scholarship policy and other information. While users browse and search for these information on the Internet, their behaviors will be recorded to form massive web log data. This huge “intention database” shows people's interests and concerns, also reflects their tendency to learn behavior in real society. And this data has the advantages of strong timeliness, high objectivity and large sample size, which provides a new perspective and opportunity for monitoring the needs of international Chinese education market.

This study firstly conducted in-depth interviews with 28 international students from 15 countries to understand the main information needs in the process of Chinese study and study abroad. Based on this, select relevant search keywords and collect these keywords from 2004 to 2016. In the search volume data of Google on these years, the final construction model empirically tests the correlation between the keyword search volume and the number of foreign students studying in China.
User Research and Keyword Selection

In order to understand the information needs and information behavior of Chinese learners better, this study conducted in-depth interviews with 28 international students in China, whom come from 15 countries and regions. Through interviews, it is found that the information behavior of Chinese learners can be divided into the following stages:

1. Motivation to learn (study abroad): At this stage, users pay more attention to China's economy, culture, history, etc. and produce keywords such as “Chinese cuisine”, “Li Xiaolong”, etc.;
2. Chinese learning and test data: generate keywords such as: "synonym", "Chinese pinyin", "HSK test", etc.;
3. Study abroad and scholarship information: generate keywords such as: “Chinese University Ranking”, “Chinese Government Scholarship”, etc.;
4. Information exchange and sharing: Chinese learners will share some Chinese learning information on social networking sites such as Facebook and Twitter, and provide new information for future international students.

Based on the above interviews, this paper selects 21 keywords as the initial keywords for research. Figure 1 shows a graph about the search traffic of keyword “近义词(synonym)” from January 2004 to December 2016 in the UK and the US, and the number of international students studying in China from UK and US. As can be seen from figure 1, the search volume of keywords and the number of students studying in China show a strong correlation and consistency of trends.

![Figure 1. Examples of Some Keywords Search Trends.](image)

Empirical Analysis

Data Sources

The number of international students studying in China used in this article is derived from the official statistics of the Ministry of Education of China. It uses data from 15 countries from 2004 to 2016. The web search data comes from Google Trends, and the keywords are collected from the web search data of 2004-2016 in the above 15 different countries.

Keyword Processing

Keyword Expansion. Considering that different users will use different search keywords for the same information needs, this study expands the initial keywords, such as expanding the initial keyword “中国文化” to: China culture; expand the initial keyword "HSK" to: HSK3, HSK4 etc. After the expansion, a total of about 1800 keywords were obtained, and this study did not consider multilingual problems, all using English keywords.

Keyword Synthesis. According to the different meanings, this paper divides the keywords obtained above into four categories and synthesizes four types of indices.

1. School class. Such keywords specifically include the names of Chinese universities, Chinese university rankings, and Chinese university majors.
(2) Study in China. In order to reduce the number of variables in the model while retaining the selected strong related keywords [9], the scholarships, Chinese cultural customs and visa keywords are all classified into the Chinese class. Scholarships mainly include school scholarships, Chinese government scholarships and scholarship applications; Chinese categories mainly include Chinese culture and customs; visas mainly include visas and visas.

(3) Chinese learning class. Such keywords mainly include synonyms, synonyms, Chinese translation, synonym dictionary and so on.

(4) Hsk class. Such keywords mainly include hsk, hskk, hsk 2, hsk 3, hsk 4 and hsk 5.

Model Building

After the above data processing, this paper obtains the data of the number of students studying in China in 13 countries as the dependent variable Y, and the corresponding search keyword indexes X1, X2, X3 and X4 are used as independent variables to construct the model verification and relationship between the Y ans X.

Unit Root Test. The data used in this paper belongs to the panel data. It is necessary to check the stability of the data before regression, and there will be pseudo-regression in the non-stationary sequence. The test results are shown in the following table. The test results of the three cases of sequence Y and X4 do not reject the null hypothesis, and the two sequences are not stable; the unit root test results of the sequences X1, X2, and X3 reject the null hypothesis, so they are all stable. sequence. Because the co-integration test is needed in the next step of this paper, the premise of cointegration test is the same order, so the original sequence needs to be differentiated. The purpose of the difference is to eliminate some fluctuations and make the data stabilize. In this paper, after performing first-order difference on all sequences, the sequence becomes stable and can be tested in the next step.

Table 1. Unit Root Test Result.

<table>
<thead>
<tr>
<th>Variable Quality</th>
<th>Statistic</th>
<th>Prob.**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>0.12901</td>
<td>0.5513</td>
</tr>
<tr>
<td>X1</td>
<td>-9.98845</td>
<td>0.0000</td>
</tr>
<tr>
<td>X2</td>
<td>-12.6641</td>
<td>0.0000</td>
</tr>
<tr>
<td>X3</td>
<td>-7.33162</td>
<td>0.0000</td>
</tr>
<tr>
<td>X4</td>
<td>3.27698</td>
<td>0.9995</td>
</tr>
</tbody>
</table>

Cointegration Test. The significance of the cointegration test is to test whether there is a stable relationship between the variables. In this paper, the Kao test in Eviews software is used to co-integrate the variables. The results are shown in Table 2. The P value of the Kao test is less than the significance level of 0.05. The assumption that “there is no cointegration relationship” is rejected, and the cointegration test is passed.

Table 2. Kao Test Result.

<table>
<thead>
<tr>
<th>Kao Residual Cointegration Test</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADF</td>
<td>-1.69153</td>
<td>0.0454</td>
</tr>
</tbody>
</table>

Figure2 shows the correlation between the respective variables and the dependent variable (the number of students studying in China). It is found that X1, X2 and X3 are negatively correlated with Y,
and X4 and Y are positively correlated, which is inconsistent with expectations. The possible reasons are as follows: (1) Those who need Chinese learning will search for information about Chinese universities, Chinese food, clothing, housing and travel, but those who search for such information may not necessarily have the intention of learning Chinese or studying abroad, and there are other intentions such as the possibility of traveling to China, so these three variables do not accurately reflect the intention of Chinese learning, and the data has noise; (2) Because this study does not consider multilingual only English keywords, and users will use their native language when searching for information, leading to English keyword search. The low amount does not accurately represent the real Chinese learning needs of countries, and the data is biased. The "HSK (Chinese Proficiency Test Pinyin Initials)" is directly related to Chinese learning, and the intentional noise is excluded. At the same time, these keywords are globally specific terms, which are not affected by the differences in the mother tongue of each country, so the language noise is not excluded. Therefore, this paper finally chooses the X4 and Y building models.

Model Establishment and Analysis

In this paper, data is used as time series cross-section (panel) data to construct a random effect model with variable coefficients:

\[ Y_{it} = 9382.52497721 + C_i + \beta_i X_{it} \]

Where \( i = 1, 2, 3..., 13 \) is the \( i \)-th country; HG (Korea), RB (Japan), TG (Thailand), etc. represent the country, in which Mongolia and Kazakhstan have finally obtained keyword data. The amount is too small to synthesize the four indices and is therefore eliminated; \( C_i \) represents the intercept term of the \( i \)-th country.

Model Analysis: In the above equation, where \( X_i \) reflects the attention of the \( i \)-th national students to hsk, 9382.52497721 is the mean of the unobservable differences, and \( C_i \) is the random difference of the \( i \)-th country, which does not change with time. From the regression results of the model, we can see that R-squared=0.953898, and the model fit is very good. F-stastic indicates the strength of the interpretation of the dependent variable by the selected variable. Prob(F-stastic) is 0.000000 < 0.05, and the model equation is significant.

Finally, this paper uses the established model to bring in the value of the number of students studying in China from the national web search data. Figure 3 shows the variation of the actual value of the number of students studying in the US and Thailand and the estimated value calculated by the model. With high prediction accuracy, the web search data of HSK keywords has certain monitoring and forecasting ability for the number of students studying in China. The search volume of hsk keywords can be used to monitor the international Chinese education market in real time.
Summary and Outlook

This paper first analyzes the relationship between network search data and Chinese education market demand based on user survey, and based on this, selects the initial keywords that can reflect the needs of Chinese education market; then expands and synthesizes keywords, and builds model validation keywords. The relationship between the search volume and the number of students studying in China, the model fit degree reached 0.945, and finally concluded that there is a long-term stable positive correlation between the HSK class keyword search volume data and the number of students studying in China, so the construction is different. The HSK keyword index in the national region can be used as a monitoring indicator for the international Chinese education market demand, and can quickly and accurately monitor the fluctuations and trends of the Chinese education market demand in different countries in real time.

However, there are some shortcomings and problems in the research process. The first is the choice of keywords. How to select more accurate and comprehensive keywords that reflect the dynamics of the Chinese education market? Secondly, the question of language, after expanding the language from English to multilingual, will the effect of the model be further improved? This is also the focus of this research.
Acknowledgement

This research was supported by Science Foundation of Beijing Language and Culture University –the Fundamental Research Funds for the Central University (19YJ040004,17PT05).

References


