A Preprocessing Method of Internet Search Data for Prediction Improvement: Application to Chinese Stock Market

Ying Liu
Graduate University of Chinese Academy of Sciences
NO.80 Zhongguancun East Road Haidian District, Beijing 100190, China

Benfu Lv
Graduate University of Chinese Academy of Sciences
NO.80 Zhongguancun East Road Haidian District, Beijing 100190, China

Geng Peng
Graduate University of Chinese Academy of Sciences
NO.80 Zhongguancun East Road Haidian District, Beijing 100190, China

Qingyu Yuan
Graduate University of Chinese Academy of Sciences
NO.80 Zhongguancun East Road Haidian District, Beijing 100190, China

ABSTRACT
The correlations between Internet search data and socio-economic Indicators have been proved in many studies, but the basis work of these studies - data preprocessing, determining the quality of the result, has lacked a systematic methodology. In this paper, we develop a comprehensive method for Internet search data preprocessing, which includes the critical steps: (a) keywords selection, (b) time difference measurement, and (c) leading index composition. Applying our method to study Chinese stock market price, we can get the leading keywords index with stable leading relation and high degree of fit. Specifically, the correlation coefficient between our leading keywords index and Shanghai Composite Index reaches 98.7%, and Granger test confirms that keywords index has significant prediction ability for Shanghai Composite Index. Adding keywords index to the AR model can reduce the MAPE from 3.8% to 1.4%, and each percentage point change of keywords index is correlated with 0.136 percentage point move in the same direction of Shanghai Composite Index in next period.

Keywords
Internet search data, Preprocessing method, Leading keywords index, Time difference measurement, Stepwise composition.

1. INTRODUCTION
Search engine, as the most general tool to get information from Internet, connects information resources and users' needs. At the same time, it also records their searching behavior. Based on hundreds of millions of search engine users' records, the Internet search data can reflect the users' concerns and intends, foreshadow their behavior trends and patterns in real lives. Therefore, using this "intentions database" to predict socio-economic indicators has become a hot topic in many studies.

These studies covered the indicators in three perspectives of microcosmic, mid-scope, and macrocosmic. From the microcosmic aspect, a good instance is epidemic symptomatic detection. Ginsberg, etc (2009) found that the percentage of influenza-like illness (ILI) search query in Google was highly correlated with the percentage of ILI physician visits, and then they built a linear model based on the search data for monitoring influenza epidemics, which can estimate the level of influenza activity 1-2 weeks ahead of the traditional surveillance systems. Following this work, Jurgen A. Doornik (2009) extended it to autoregression model with calendar effects, and improved the prediction accuracy. In other empirical analysis, it also got good results of using Internet search data for forecasting movie box office, the popularity of online games and songs, the traffic of website and so on (Sharad, 2009; Heather, 2010).

From the mid-scope aspect, Choi and Varian (2009) did empirical test on the sales of U.S. automobile, housing, travel and other industries, they put keywords search frequency as a new independent variable adding to the traditional time-series models, and found that the prediction accuracy of above industries was significantly improved. Lynn and Erik (2009) also made study on the U.S. housing market and found that the search data had strong predictive power on the sales and prices of U.S. housing. From macrocosmic aspect, current studies mainly focused on the forecast of unemployment rate and private consumption. Askitas and Zimmerman(2009) demonstrated there existed strong correlations between keyword searches and unemployment rate based on monthly German data, and found significant improvements in prediction accuracy by using Google Trends. The similar work had done by Suhoy (2009) and Choi & Varian (2009) to study unemployment rate of Israel and US. As for consumption area, Kholodilin etc (2009) compared the growth rates of the real US private consumption based on both the conventional consumer confidence indicators and the Google indicators, and the results showed that the latter were 20% more accurate than the former. The predictive power of Internet search data was also confirmed by Torsten (2009), Nicolás(2009), and Marta(2009).

The above studies make some progresses in empirical analysis, but the empirical results are highly dependent on the method of search data processing. Especially in Chinese search market, there are few effective search keywords recommended by search engine. Consequently, it would get apparently different results when selecting different keywords, or making different processing method on the search data. This proves the data preprocessing method plays a critical role in the quality of the empirical study. However, current researches haven't yet formed a complete methodology on the Internet search data preprocessing. For example, by what standards or principles are keywords selected? How to measure the time-difference relationships between keywords and target indicator? How to composite a leading keywords index to reflect the target indicator trend? These questions have not yet answered completely by current literatures.

Taking these questions as a starting point, this paper will systematically introduce a preprocessing method of Internet

search data based on Chinese search data and Chinese stock market. Our method is not only applicable to current prediction, but also probably to predict the future trends of some economic or business indicators.

The structure of the paper is as follows: in Section 2 we introduce the definitions of some concepts related to Internet search data, and the source of Chinese Internet search data; Section 3 mainly demonstrates the whole flow of our preprocessing method; Section 4 refers to the details of the keywords selection; Section 5 is about the method of time difference measurement; Section 6 is about the method of leading keywords index composition; In Section 7 we test the predictive power of the leading keywords index; Section 8 is our conclusion and future work.

2. Internet Search Data and Related Concepts

2.1 Related Concepts and Definitions

Before introducing the data preprocessing method, we first define and explain following related concepts.

Search engine: a system for searching information on Internet and returning search results under users' queries (keywords). The most commonly used English search engine is Google, and current literatures mostly based on Google data. While Baidu is the most commonly used search engine in Chinese market, so the data of this paper comes from Baidu.

Search keyword: also called search query, is the text that a user enters into Internet search engine to satisfy his or her information needs.

Search volume: also known as search term frequency, or search attention, the number or frequency that a certain keyword is queried by users within a certain time.

Keywords index: generally, one keyword only reflects one side of an event according to a certain users' perceptions, and measuring the vast majority of users' expectations of an event needs the keywords as full as possible. In order to do this, we composite the keywords related to target indicator into a keywords index.

Shanghai Composite Index: is an index of all stocks (A shares) that are traded at the Shanghai Stock Exchange, and its base day is December 30, 2005. We choose Shanghai composite index as target indicator to reflect the market's overall trend.

2.2 Data Source

The search data in former literatures almost come from Google trends (www.google.com/trends), or Google Insights (www.google.com/insights/search). While Baidu holds 70% market share in Chinese search market, so the data from Baidu is more universal for Chinese market. Our data resource comes from Baidu Zhishu (zhishu.baidu.com), which can provide search volumes of certain keywords from June 2006 till 2010.

Conference'10, Month 1–2, 2010, City, State, Country.

3. The Whole Flow of Data Preprocessing

The purpose of data preprocessing is to processing the raw data into the data format which can be used directly to do empirical test and prediction. The whole flow of our data preprocessing is shown as Figure 1. The first step is to select the keywords which are related to the target indicator, and then obtain the search volume data of each keyword from Baidu Zhishu. After necessary data filtering, we measure the time difference between each keyword and target indicator. In this paper, we use two methods (cross-relation analysis and K-L distance) to calculate time difference separately, and choose the keywords those present leading relation under both two methods. Next, we need to test the stability of the leading period using spectrum analysis, and remove the unstable keywords, then, the remained keywords would have stable leading period with target indicator. In the end, we composite all the stable leading keywords into an index according to the leading period and weight, and this index can be used as an important variable for prediction model. Each step in the whole flow will be described in next section.

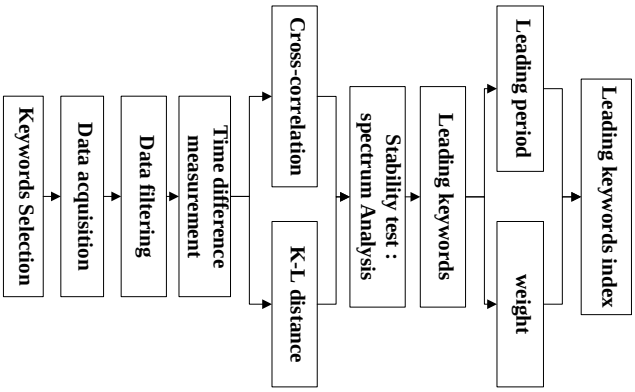

Figure 1: the whole flow of search data preprocessing

4. Keywords Selection and Search Data Acquisition

4.1 Keywords Selection

Keywords selection is the basis of data preprocessing. As the heterogeneity of users, they have different points of view towards one object, so the searching keywords are also characterized by diversity. It should be a set that contains the most common keywords related to the target indicator. But the keywords are not the more the better, because when the coverage of keywords reaches a certain level, then the marginal contribution of adding new keywords is very limited (Hulth, 2009), while the cost grow rapidly. For example, Ginsberg, etc, select 45 core keywords from 50 million keywords with the method of exhaustion through computers and this method is limited by the resources and hard to copy. The keywords existing in current literatures often rely on the experience or directly refer to keywords recommended by Google. The weakness of experience is too subjective and easy to produce omission; the keywords recommended by Google are broad, but are not relevant to specific subjects. So, keywords recommended by the search engines can only be an alternative method, besides, we also use text mining of forums and news to improve the user coverage.

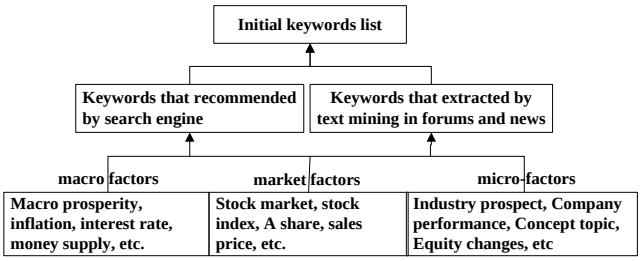

Figure 2: method for keyword selection

Here we gives an example based on "Shanghai Composite Index" to illustrate the specific steps of keywords selection:

(1) Analysis of influencing factors. First, summarize and class the factors that influence the stock market; then, through the study of historical literatures, the factors in this paper will be divided into three levels: the macro-level, the market level and the micro level. The search query of these factors on behalf of the need or expect of users, are also reflections of changes in the stock market.

(2) Obtaining keywords through search engines. On the basis of analysis of influencing factors, we can firstly obtain the alternative keywords through search engines. Specifically, we can refer to the keywords recommendation address of Baidu:support.baidu.com/topic/18.html; the keywords recommendation address of Google: www.google.com/sktool.

(3) Internet text collection and keywords extraction. We collect the text of the three types of influencing factors from Internet news, forums and blog through the search tools of press search, forums search and blog search, entering 3-5 basic keywords into the tools, and then save the sub-period results in text format. Make text mining on Internet text with Chinese word segmentation tool, and extract a list of keywords relevant to the subject with high frequency. After merging the list and the ones recommended by search engines, we get a set of initial keywords in this paper. Besides the above process, the following principles must be considered:

Important economic meaning: the selected keywords can represent a side of impact factors of the target indicator and reflect some behaviors of users, the combination of all the keywords can reflect the main aspects of the target indicator.

Sufficient statistical meaning: the keywords selected should be some continuity of data series, easier to measure the stable relationship with the target indicator.

Corresponsive to target indicator: there is a strong correspondence between the data series of keywords and the wave of target indicator in peaks and valleys, so as to fix the leading indicator.

After the above steps, we get the initial keywords with the number of 2103 for the following data preprocessing.

4.2 Search Data Acquisition and Filter

According to the keywords list, we enter each keyword into Baidu Index (zhishu.baidu.com) then we can view the amount of time series data of each keyword. But Baidu index does not provide downloads of the search data. In period to search for convenient access to the historical data, we make a Java-based spider program for the download of time series data and the raw data format captured from Baidu index is shown in Table 2. There are few keywords whose search volume is zero in the original data. Because the frequency of such keywords is too low, Baidu index cannot show their search volume. Such keywords are called invalid keywords and removed from the original set and the number of the remaining valid keywords is 1926.

5. Time Difference Measurement

Firstly, taking two keywords ("GuPiaoRuMen", "ShangZhengZhiShu") as an example, we compare the relationship between the search volume and the closing price of Shanghai Composite Index, the comparison figure shows that the search curve of "GuPiaoRuMen " is ahead of Shanghai Composite Index price movements, and the search curve of "ShangZhengZhiShu" is lagging behind the Shanghai index price movements.

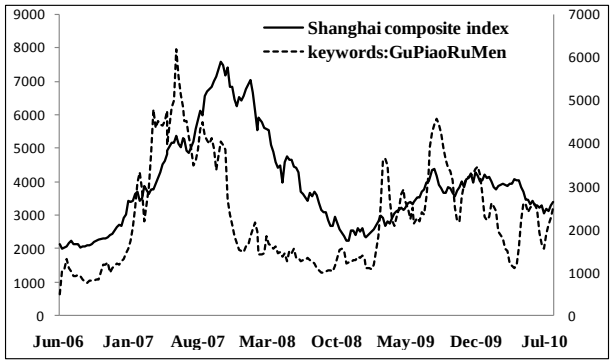

Figure 3: comparison Shang composite index with the keywords "GuPiaoRuMen"

Figure 4: comparison Shang composite index with the keywords "ShangZhengZhiShu"

Obviously, different keywords show different cross-correlation and we call the keywords with leading characteristics the leading indicators, which are the leading roles of the changes on the stock market. Only these keywords have the predictive value. The judgment and filter rules will be introduced in this section in detail.

We give two parameters to each word: leading period and correlation coefficient. The leading period is to measure the leading character of the keywords to target indicator. If the leading period is less than zero, it presents that the keywords is leading to the target indicator; if equals to zero, it presents coincident relationship; if more than zero, lag relationship. The correlation coefficient means the similarity between the cure of keywords and the cure of Shanghai Composite Index, the bigger

the coefficient is， the more similar between them. We adopt two methods to calculate the leading period of each keyword in this paper, and select the most predictive keywords with the use of the two parameters.

5.1 Time Difference Relevance

The time difference relevance is a common method to verify the leading, coincident or lagging relations of economic time series. The formula is as follows.

$$r_l = \frac{\sum_{t=1}^{n}(x_{t+l}-\bar{x})(y_t-\bar{y})}{\sqrt{\sum_{t=1}^{n}(x_{t+l}-\bar{x})^2\sum_{t=1}^{n}(y_t-\bar{y})^2}}, l = 0, \pm1, \pm2, \cdots, \pm \mathrm{L}$$

In the above formula, r_l denotes the coefficient of the cross-correlation with l; y denotes year yield of Shanghai Composite Index, $\bar{y}$ as the mean; x denotes the annual rate of change of keywords search, $\bar{x}$ as the mean; l is the leading period to x. We define the max time difference period as the leading order, and this cross-correlation presents the relationship between the two.

5.2 K-L Distance

K-L distance is an indicator proposed by Kullback and Leibler to determine the proximity of the two probabilities distribution, the closer the value is to 0, indicating the closer the selected indicators to the target indicator. In actual calculation process, given the target indicator: $y = \{y_1, y_2, \mathrm{K}, y_n\}$, standardize it and make the sum of the indicator as a unit 1, the processed sequence denoted by p:

$$p_t = y_t / \sum_{j=1}^{n} y_j, t = 1, 2, \mathrm{K}\ n$$

Given selected indicator: $x = \{x_1, x_2, \mathrm{K}, x_n\}$, standardize it, denoted by q:

$$q_t = y_t / \sum_{j=1}^{n} y_j, t = 1, 2, \mathrm{K}\ n$$

The formula of K-L distance is as follow:

$$K_l = \sum_{i=1}^{n} p_i (p_i / q_{i+l}), l = \pm1, \pm2, \mathrm{K} \pm L$$

$$K_{l^*} = Min(K_l), l = \pm1, \pm2, \mathrm{K} \pm L$$

Separately calculate the K-L distance for 2L+1 times, select the minimum value as K_{l^*} that is the K-L distance of selected indicator toward target indicator. l_* presents the leading order; l_* <0 indicates a leading and l_* >0 for a lag.

We calculate the cross-correlation between the selected keywords and target indicator with the two methods separately. When the leading orders calculated through the two methods are both less than zero and the distance between them is small, we award the keywords have the leading character. After abvoe filter preprocess, 378 keywords remain.

5.3 Stability Test for Leading Relations: cross-spectral analysis

In general, the time difference relevance and K-L distance can measure the leading character of keywords. We cannot study the dynamic change of leadership of different periods with the two methods. However in practice, the time difference between certain keywords does have unstable characteristics. As the specific performance that at a different time period leading to different orders, some keywords even transfer its leading character in early stage into lag in later stage, thus, the stability of the keywords needs to further determine. Cross-spectral analysis is used to do the stability test in this paper.

According to cross-spectral analysis, the time-series is made up of independent components which have different amplitude, frequency and phase of cycle. Transforming the time series from time domain to frequency domain by Fourier transform, so as to decompose each frequency component and compare periodic change of each component. The cross-spectral function can be used to test the dynamic changes of leading relationship, follow these steps:

First, remove the trends among the target indicator and alternative indicator using HP filter, getting the smooth sequence: $\{x_t\}$、$\{y_t\}$. According to the spectral analysis theory, both the cross-spectral density functions are the Fourier transform of cross-correlation function, that is,

$$f_{x,y}(\omega) = \int_{-\infty}^{\infty} r_{\mathrm{x,y}}(\mathrm{t}) \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{dt} = p_{\mathrm{xy}}(\omega) - \mathrm{i} q_{\mathrm{xy}}(\omega)$$

In the above formula, $r_{\mathrm{x,y}}(\mathrm{t})$ denotes cross-correlation function; ω denotes frequency. Function value is generally complex, the real part $p_{\mathrm{xy}}(\omega)$ is called co-spectrum and the imaginary part $q_{\mathrm{xy}}(\omega)$ is called quadrature spectrum. In practical terms, they are usually translated into the coherent spectrum and phase spectrum to analyze. The coherent spectrum is a standardized mean of amplitude product between the two series with frequency w, which can measure the absolute correlation of them. The closer its value is to 1, the stronger correlation is. The phase spectrum is a mean that reflects the phase shift between the two series with the frequency w. If its value is negative, leading relationship is detected; otherwise, there exists lag relationship. The formulas are as follows.

Coherent:

$$\mathrm{c}_{xy}(\omega) = \frac{\left|\mathrm{f}_{\mathrm{xy}}(\omega)\right|}{\sqrt{\mathrm{f}_{\mathrm{x}}(\omega)\mathrm{f}_{\mathrm{y}}(\omega)}}, 0 \le \mathrm{c}_{xy}(\omega) \le 1$$

Spectrum:

$$\phi_{xy}(\omega) = \tan^{-1}\left(-\frac{\mathrm{q}_{\mathrm{xy}}(\omega)}{\mathrm{c}_{\mathrm{xy}}(\omega)}\right)$$

The correlation and leading character between alternative indicators and target indicators with different frequency (or different period) can be tested by the coherent spectrum and phase spectrum. The selected keywords show consistent leadership and correlation in major circle, and finally we get 190 keywords.

6. Leading Index Composition

There are obvious differences in relationship between the search data cure of different keywords and Shanghai Composite Index, indicating one of the keywords can only reflect one kind of the behavioral characteristics of a stock market, the overall trend of the market must be described by many keywords together. Therefore, we should composite the leading keywords into leading index. The goal of index composite is to maintain the leading role to the target indicator, as well as the highest possible correlation. To achieve the goal, there are two key steps in the course of synthesis: the synthetic weights and the rule of keywords combination. The following are the introductions of the two areas.

6.1 Composition Weights

There are two main methods to determine the weights in the study of economic boom: one is method of systematic assessment (Moore,Shiskin， 1967； Boehm， 2001) -rate the selected indicator according to the principle of prior evaluation and then make the rates as the weights. These rating principles include the rationality of the economic implications, statistical adequacy, the stability of peak and valley and smoothness of these sequences. The method is comprehensive but highly subjective. The other method is empowering according the correlation coefficient (Zhu Jun, Wang Changsheng,1993; Ginsberg,etc， 2009). In this method, the greater the correlation between selected indicators and target indicators, the greater the weight is. It has a certain objectivity, but sometimes there are poor cases.

We combine the two methods in this paper. The weights are defined as a function of correlation coefficient, that is w=f(r), which means the weight of some keywords is not only decided by the similarity (not entirely decided), but also affected by other principles which will be given by the method of AHP. Finally, we combine the two methods to determine the weights.

6.2 Composition Rules of Keywords

The leading period and weights should be considered in the course of the combination of keywords. Firstly, each keywords sequence multiplied by corresponding weight, and then adjust the time difference, make each keyword sequence in accordance with the target indicators with staggered alignment according the leading order. Next, judge keywords which can be better fit the target indicator. This paper uses partial F test to decide whether to add some keywords into the index, makes the synthesis of keywords with the thinking of stepwise regression (Stepwise), so this method is called stepwise composition method. Its workflow is shown in Figure 5.

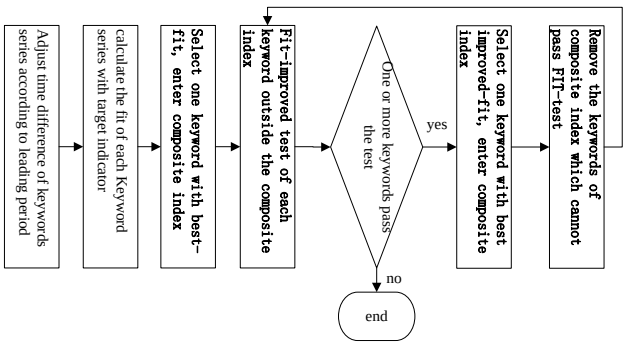

Figure 5: the flow of stepwise composite method

6.2.1 Fit-Improved Test

Partial F test can be used to study whether the goodness of fit is improved after adding certain keywords to the composite index, as follows:

Model 1: $y = \beta_0 + \beta_1 Index_{j-1} + \varepsilon$

Model 2: $y = \beta_0^{'} + \beta_1^{'} Index_j + \varepsilon^{'}$

In the above formulas, y denotes target indicator; β_0、β_1、ε separately denote the intercept, coefficient and random interference terms; $Index_j$ indicates the index made up of j keywords, the formula is $Index_j = \sum_{i=1}^{j} w_i x_i^{li}$, w_i is the weight of the i keyword, x_i^{li} presents the sequence after alignment by the leading period of the i keyword.

When j>1, given Fitj-1 as the goodness of fit of model 1, then Fitj as the goodness of fit of model 2.

Given: $\Delta Fit_j = Fit_j - Fit_{j-1}$

The original hypothesis H0: ∆ Fitj≤0, meaning that it cannot improve the goodness of fit when adding the i keyword into the index.

The alternative hypothesis H1: ∆ Fitj>0，meaning that it can improve the goodness of fit when adding the i keyword into the index.

Statistics: $\dfrac{\Delta Fit_j / 1}{(1 - Fit_{j-1}) / n - 2} \sim F(1, n-2)$

If the statistics are significant, it means that the keyword should be added to the composite index, and vice versa.

6.2.2 Stepwise Composite Method

In a multi-regression, stepwise regression method is used to select significant independent variables. Here we apply it to the composite of the index in this paper, in order to select keywords which can improve the goodness of fit significantly, and the processes are shown in figure 5.

a. build regression model between the target indicator and each keyword, select the one which can achieve the max goodness of fit.

b. make test of improving goodness of fit on the keywords outside the synthetic list separately, once certain keywords passes the test, then select the keyword which achieves the most important influence of the improving goodness of fit into the composite index.

c. make the test of goodness of fit to the other keywords in the synthetic list, remove the ones that cannot pass the test.

d. repeat b, c steps, till no keywords outside the synthetic list can improve the test by fitting, and then end the cycle.

After these steps, the leading indicators which have the biggest goodness of fit of the target indicator can be got.

7. Empirical Tests on the Keywords Index and Shanghai Composite Index

Comparing the keywords index (Indext-1) with one period ahead and Shanghai Composite Index, we find that there is strong consistency of peaks, valleys and the trend of fluctuations of the two, with the correlation coefficient up to 98.7%. To test the predictive performance of the keywords index for Shanghai Composite Index, first, we use Granger causality test in this section, and then check the improved predictive ability of keywords Index for Shanghai Composite Index through the regression model.

Granger causality test can exam whether the predictive ability exists between the variables. The specific method is as follows. First we estimate the explained degree of y by its own lag, and then verify whether the introduction of x can improve the interpreted level of y. The original assumption is that x cannot improve the interpreted level of y, if reject the original hypothesis, then we can say x can Granger cause y. Testing the level-values and difference values of Indext-1 and yt respectively, results are shown as follows.

Table 1: Granger test results of keywords index and Shanghai Composite Index

	null hypothesis	samples	P-values
Level-values	Index cannot Granger cause y	189	0.00
	Y cannot Granger cause Index	189	0.23
first difference	ΔIndex cannot Granger cause Δy	188	0.00
	Δy cannot Granger cause ΔIndex	188	0.02

The results show that, either the level-values or the difference values, the keywords index can Granger causes Shanghai Composite Index significantly at the 1% level; while Shanghai Composite Index cannot Granger cause the keywords index at the level-values, but after the differential treatment it can set up significantly at the 5% level. These indicate that the synthesis of the keywords index in this paper has significant predictive ability for Shanghai Composite Index; conversely, Shanghai Composite Index cannot predict the keywords index. The fluctuations of them are the mutual influence and the keywords index volatility can Granger cause the fluctuations of Shanghai Composite Index more significantly.

Two models are built in this chapter, model ① is the self-regression model of Shanghai Composite Index and then we add the keywords index into model ②:

$$\log(y_t) = \beta_0 + \beta_1 \log(y_{t-1}) + \varepsilon \qquad ①$$

$$\log(y_t) = \beta_0 + \beta_1 \log(y_{t-1}) + \beta_2 \log(Index_{t-1}) + \varepsilon \qquad ②$$

The results of the empirical test show that model② can fit better than model①, and the average error rate of 1.4% of model② is significantly better than the 3.8% of model①. Log(indext-1) passing the test at 1% level. The coefficient 0.36 indicates that each percentage point change of keywords index is correlated with 0.136 percentage point move in the same direction of Shanghai Composite Index in next period.

Table2: fitting results of keywords index and Shanghai Composite Index

	model①	model②
Adjust R^2	0.974	0.978
intercept	0.160^{*}	0.212^{***}
$\log(y_{t-1})$	0.980^{***}	0.738^{***}
$\log(index_{t-1})$	——	0.136^{***}
MAPE	0.038	0.014

Note: ***represents at 1% level; *represents at 10% level.

Figure 6: comparison actual Shanghai composite index with the fitted value

8. Conclusion and Future work

The relationship between the Internet search data and the socio-economic behaviors is a hot research topic in the past two years abroad and the preprocessing of the search data is the basis for the study. However, it is lack of a systematic methodology for the problem in the current study. In this paper, we develop a comprehensive method for Internet search data preprocessing, which include the critical steps: (a) keywords selection, (b) time difference measurement, and (c) leading index composition. There are three main innovations in this paper:

Firstly, in the selection of keywords, we can acquire as full as possible keywords with the smallest cost through the Chinese process and principles of keywords selection based on the influence factors analysis and Internet text mining.

Secondly, in the relationship judgment of the keywords and target indicator, the time difference relevance, K-L distance and cross-spectral analysis are adopted to judge the leading character and stability of the keywords in this paper. Then the keywords with predictive power can be got.

Thirdly, in the leading keywords composition, we use stepwise composition method to get better goodness of fit between leading indicator and target indicator.

The keywords index can have the character of leading and wonderful goodness of fit after the preprocessing method, and can be used as a better data base for the empirical analysis and prediction study. The predictive power does exist between the leading keywords index and target indicator confirmed by the Granger test and the regression results show that every 1 percentage point change of keywords index, Shanghai Composite

Index of the latter period moves 0.136 percentage point in the same direction.

There are also some shortages in this paper, although the Internet search data has universality, it isn't the only channel to get information. Therefore, our future work include how to combine search data with other Internet data (such as Internet browsing, online comments, etc.), and how to combine Internet data with traditional data.

9. ACKNOWLEDGMENTS

This research is supported by the National Natural Science Foundation of China under Grant 70972104, the National Natural Science Foundation of China under Grant 71172199, Foundation of Dean of Graduate University of Chinese Academy of Sciences under Grant Y15101QY00, and Postdoctoral Science Foundation under Grant 2011M500422.

10. REFERENCES

[1]. Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant, 2009, Detecting influenza epidemics using search engine query data [J], Nature 457, 1012~1014.

[2]. Jurgen A. Doornik, 2009, Improving the Timeliness of Data on Influenza-like Illnesses using Google Search Data [J], Working paper.

[3]. S Goel, JM Hofman, S Lahaie, DM Pennock, DJ Watts, What Can Search Predict，workingpaper，2009.

[4]. Heather.LR Tierney, B Pan，A Poisson Regression Examination of the Relationship between Website Traffic and Search Engine Queries，workingpaper，2010.

[5]. H Choi, H Varian .Predicting the Present with Google Trends,workingpaper.Technical Report, Google Inc,2009.

[6]. H Choi, H Varian , 2009,Predicting Initial Claims for Unemployment Benefits.

[7]. Lynn Wu ， Erik Brynjolfsson ， The Future of Prediction—how google searches foreshadow housing prices and sales，12，2009.Workingpaper.

[8]. Tanya Suhoy.Query Indices and a 2008 Downturn: Israeli Data. Bank of Israel, 2009.

[9]. N Askitas, KF Zimmermann. Google Econometrics and Unemployment Forecasting[J].Applied Economics Quarterly, 2009.

[10].Konstantin A. Kholodilin.Maximilian Podstawski. Boriss Siliverstovs.Constantin B ürgi. Google searches as a means of improving the nowcasts of key macroeconomic Variables, Discussion Papers. Berlin, November 2009

[11].Torsten Schmidt, and Simeon Vosen.Forecasting Private Consumption: Survey-based Indicators vs. Google Trends[J]. Technische Universität Dortmund, Department of Economic and Social Sciences,2009.

[12].Nicolás Della Penna and Haifang Huang,2009, Constructing a Consumer Confidence Index for the US Using Web Search Volume, workingpaper,

[13].Marta. Consumption and Information: An Exploration of Theories of Consumer Behavior using Daily Data.workingpaper,2009.

[14].Hulth A, Rydevik G, Linde A. Web queries as a source for syndromic surveillance[J]. PLoS ONE 2009 Feb 6;4(2):e4378

[15].Dong Wenquan, Gao Tiemei, Business Cycle Analysis and Prediction Analysis, Jilin University Press, 1998

[16].Moore, G.H., and Shiskin, J, 1967, Indicators of Business Expansions and Contractions [M], NBER Occasional Paper No. 103.

[17].Boehm, E.A, 2001, The Contribution of Economic Indicator Analysis to Understanding and Forecasting Business Cycles [J], Indian Economic Review, 36, 1~36.

[18].Zhu Jun, Wang Changsheng, Economic boom of theoretical methods of early warning systems [M], China Planning Press, 1993

[19].Dong Zhiqing, Wanglinhui, China's stock market and macroeconomic volatility Relevance: Based on Wavelet and cross-spectral analysis of the comparative test, financial research, 2008.8, PP :39-52.

[20].S. Kullback and R. A. Leibler. On information and sufficiency. Ann. Math. Stat., 22:79–86,1951.

[21].Allan, D. W.; Daams, H. Picosecond time difference measurement system. Electronic Industries Association, 1975, p. 404-411.