ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 14, No. 1, 2019, *pp*.044-051

Study on Prediction Model of Personal Economic Level Based on Text Analysis Using Chinese Classified Lexicon

Yahui Chen, Zhan Wen, Xia Zu, Yuwen Pan and Wenzao Li*
School of Communication Engineering, Chengdu University of Information Technology, Chengdu, Sichuan, 610225, China
E-mail: chenyahuicyh@gmail.com
(Received January 09, 2019, accepted March 18, 2019)

Abstract. Obtaining economic situation of the group is a key step in understanding the socio-economic situation like the division of the rich and the poor. But the traditional way to obtain economic situation of the group is based on the survey data of professionals and mathematical models. Such methods are time-consuming and too dependent on professionals. Therefore, the use of data mining techniques to judge and predict the economic situation of the group came into being. Such methods are efficient that can overcome the shortcomings of the traditional methods. In this paper, we started by acquiring the individual's economic level and finally established a personal economic level prediction model. Through large-scale access to the individual's economic level, the economic level of the group can be obtained. We analyzed the Chinese text data published on the network by Individuals with logistic regression model to explore whether the above text data can reflect a person's economic status. The experimental results indicate that personal created textual data is able to forecast the individual's economic level accurately and certain categories of vocabulary have an impact on the individual's economic level.

Keywords: text analysis, logistic regression, personal economic level, prediction model.

1. Introduction

The acquisition of the group's economic level is of great significance for understanding the socio-economic situation such as the macroeconomic development of a certain country or region and the division of the rich and the poor, etc. But the traditional methods rely on the investigation and analysis of professionals. It often takes a long time and usually lags behind the government's macro-control policy formulation or business decision-making needs. Therefore, how to get the group’s economic level efficiently is a key issue for the government, enterprises, and scientific research institutions. With the development of computer technology, many researches based on economic data combined with machine learning to analyze the socio-economic situation.

However, such research usually faces difficulties in data collection because economic data like fiscal revenue and historical GDP are often not open to the public and owned by government agencies and large enterprises. At the same time, the researchers found that in addition to economic data are related to the economic level of the group, Web search data, text data in Internet forums and blogs can also predict the socio-economic situation. E.g., Choi H, Varian H forecast the unemployment rate trends in Germany, Israel, Turkey, Italy, and the United States through research on employment entry vocabulary and recruitment query index in Google Trends, achieved better results than the national unemployment rate prediction model based on professional forecasters survey[1]; Todd H. Kuethe et al., used a series of reports from the US Department of Agriculture and archived data on agricultural income from the US Department of Agriculture website to forecast US agricultural net income and achieved good results[2]. Therefore, using the public text data on the network can predict the socio-economic situation such as macroeconomic growth, personal consumption level, and group economic income level.

This study is based on the web text data published by individuals, such as autobiographies, Personal blog in Sina company, transcripts of interviews and speeches, literary works, etc. With the goal of acquiring the economic level of the group, we started with the establishment of a personal economic level prediction model, i.e. exploring whether the personally created text can reflect a person's economic status. In the study, we used logistic regression model to fit and predict the data. Finally, we selected the best model’ parameters that can accurately identify the individual's economic level by comparing the minimum mean square error(MMSE) of the predicted results. And we divided the economic status of the population into two

categories. One is the extremely affluent population, i.e. rich people with a large number of wealthy records, such as members on the Forbes rankings over the years and Chinese Fortune 500 list. The other is the population of the general economic level, that is, the non-extremely affluent population whose economic level is above the poverty line. The level of the groups' economy can be obtained by forecasting a large number of individuals' economic levels.

This paper has three main contributions: First, there is currently no use of online public text data to discriminate the group's economic level. This kind of data is easy to obtain, and does not require people with relevant professional background knowledge to analyze. This can improve research efficiency and save costs; Second, in the research process, Tsinghua University's word segmentation vocabulary is used to classify keywords and convert text data into vectorized data that can be used for training. This is a new method to text data vectorization; Third, our study has improved the single point of application of previous research results. Through the establishment of the individual economic level forecasting model, the group economic situation can reflect the gap between the rich and the poor in a certain region, which can optimize the implementation of poverty alleviation economic policies and improving people's livelihood.

The rest of the paper is organized as follows: Sect.2 Introduces domestic and international research on topics similar to this study and detail the theoretical knowledge of the research methods. Section 3 presents the detailed process of the experiment and the presentation and analysis of the results in each of the experiments. The paper is concluded in Sect.4. The five section presents the shortcomings of the experiment and puts forward suggestions for the subsequent improvement work.

2. Related Works

This section will introduce the related work done by the predecessors in the prediction of the socio-economic situation. They are the traditional forecasting methods based on statistics and the methods of using machine learning to analyze different kinds of data on the network to establish a predictive model. And related models and methods used in related works will be introduced too.

Research on predicting socio-economic situation based on network data

The prediction of the socio-economic situation began in the early 20th century. The traditional method of socio-economic forecasting is based on mathematical theory such as economics and statistics combined with historical economic data to obtain prediction results. In the forecasting process, the amount of manual work is huge, it takes a long time, and often has hysteresis. The researchers' misjudgment in several recessions by using the traditional methods in the US economy applied the above problems. Therefore, the current research on machine learning to predict social and economic conditions with Internet data has gradually become a hot topic. The data used in the prediction of the socio-economic situation generally fall into two categories: The first category is economic data, such as GDP value, per capita income, and regional economic growth data and so on; The other type is web text data, such as Google or Baidu search annual keywords, annual economic analysis reports.

Foreign researchers such as David E. Bloom used the economic growth data of various countries between 1980 and 2000 and the proportion of working-age population, combined with the economic growth model to predict the macroeconomic growth of each country from 2000 to 2020[3]; Hsiao-Tien Pao used Taiwan's electricity consumption data and Taiwan's historical GDP values to explore whether there is a predictive relationship between Taiwan's economic growth and historical power consumption data. He effectively proved that there is a correlation between Taiwan's electricity consumption data and Taiwan's economic growth[4].

However, in the course of research, since some economic data is not disclosed, it is often difficult to obtain experimental data. Therefore, researchers are not constrained by the use of economic data to predict the socio-economic situation, but to expand the scope of research data to network text data that is easier to obtain. Su Z found that a series of employment-related macro indices can be predicted by using the "unemployment" related keyword search index in Baidu Index and Google Trends[5]; In 2012, the United Nations launched the Global Pulse project to forecast the socio-economic situation of rising unemployment and rising prices in poor countries based on data such as social networks, and to adjust humanitarian assistance project policies through digital early warning signals to help More areas to get rid of poverty more effectively[6].

According to the above analysis of relevant research at home and abroad, it can be found that foreign researchers have begun to study social economic forecasting earlier than domestically, and have made many

breakthrough conclusions, which provide a lot of ideas and methods for follow-up research in China. Although China's research on social and economic situation prediction is later than foreign countries, it has also achieved rich results in recent years with the popularity of the Internet and the arrival of the data age. In the study of predicting the socio-economic situation based on economic data, researchers at home and abroad have achieved better results than the prediction of socio-economic situation based on network text data, so the development of research on network text data using text analysis technology is still in progress.

Principal component analysis(PCA)

Principal component analysis(PCA) is one of the common methods for dimension reduction of data features. It replaces the original features with a smaller number of new features. The new features are linear combinations of old features. These linear combinations maximize sample variance and try to make the new features uncorrelated. This makes it easy to study and analyze the impact of each feature on the predictive results of models , and effectively reduce the complexity of models and increase the speed of the operation. It is generally believed that the final cumulative contribution rate of features above 85% means that most of the information is included.

K-fold cross-validation

Cross Validation is a statistical analysis method that can be used to verify the performance of a classifier. The basic idea is to group the raw data (dataset), one part as a training set and the other as a verification set (validation set). First, the classifier is trained with the training set, and then the training set is tested by the verification set. There are three common cross-validation methods: Hold-Out Method, K-fold cross-validation, Leave-One-Out Cross Validation. In the study, the second method is used, namely K-fold cross-validation.

K-fold cross-validation divides the raw data into K groups equally, and each set of data is used as a verification set, and the remaining K-1 data is used as a training set. The average of the obtained K classification accuracy rates is used as a performance indicator of the classifier. K-fold cross-validation can effectively avoid over-learning and under-learning, and the results obtained are more persuasive. Therefore, K-fold cross-validation is also chosen as a measure of classifier performance in this experiment.

Logistic regression

Logistic regression is a widely used generalized linear regression model that can be used in binary and multiple regression applications, e.g. automatic disease diagnosis and economic forecasting. Although the logistic regression model has a general effect, the final fitted parameters represent the influence of each feature on the results, which facilitates the clear analysis of the specific impact of the selected features on the individual economic level.

In the binary regression's fitting process of this paper, we used the average gradient descent algorithm as the loss optimization algorithm to calculate the maximum likelihood probability. To prevent over-fitting of the model, an L2 regular penalty was added to the loss function. During the experiment, the model with the MMSE were got by adjusting the coefficient of regular penalty term and the number of packets in the K-fold cross-validation.

Let the argument input to the model be X, It is composed of multiple input variables just like x1,x2.... Each variable represents a feature selected in the experiment, And β , which represents the input weight, is the importance of each feature and is the amount that needs to be solved. The logistic regression mapping function fixes the model output Y between [0, 1], narrowing the range of the output, which is convenient for analyzing the category of the judgment result. Mapping function's formula of logistic regression model is as follows:

$$y = \frac{1}{1+e^{-(\beta X+\theta)}} \quad (1)$$

In the actual experiment, even if there is no input, the model will produce a certain output due to the noise data, so θ is the inevitable bias.

THU Open Chinese Lexicon (THUOCL)

THUOCL (THU Open Chinese Lexicon) is a set of high-quality Chinese vocabulary compiled by the Natural Language Processing and Social Humanities Computing Laboratory of Tsinghua University[7]. The vocabulary comes from the social labels of mainstream websites, search hot words, input method lexicon,

DOI https://doi.org/2024-JICS-22431 | Generated on 2025-04-05 21:21:36

etc. The lexicon covers a total of 157,173 entries in 11 categories including information technology, finance, idioms, place names, historical celebrities, poetry, medicine, diet, law, automobiles, and animals. The terminology has many entries and is accurate. It can be used in topics such as Chinese automatic word segmentation to improve the effect of Chinese word segmentation.

3. Research on Personal Economic Level Forecasting Model Based on Network Text Data

Introduction to the experimental process

Based on the established process of machine learning training data, the experimental process of this paper is as follows.

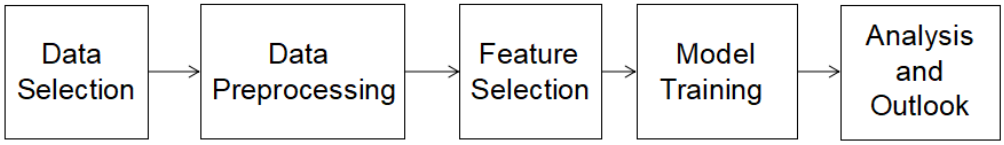

Fig. 1: experimental process

1) Data collection: Text data published by different economic level's populations on the Internet were collected and screened in this step.
2) Feature selection: The classification of words in THU Open Chinese Lexicon was taken as the feature in the experiment, and the principal component analysis(PCA) method was used to select the final required features.
3) Data preprocessing: The obtained data were vectorized in the step, i.e. the data were segmented; the keywords were extracted from each text data; the word frequency statistics based on the categorical lexicon were processed. The vectorized data would be imported into machine learning models for training.
4) Model training: The vectorized data obtained in the previous step was input into the selected machine learning model in this part. And the mean square error(MSE) of the output results was used as a measurement standard of selecting the best parameter.
5) Analysis and outlook: Based on the results obtained after the model training, we carried out the analysis and drawn the conclusion. Also, we analyzed the deficiencies in the experiment and put forward suggestions for improvement.

Data collection and feature selection

In the study, we used network texts published by the population at different economic levels as the data foundation. The collected text data is mainly divided into the following two categories.

The text of extremely affluent populations. The extremely affluent population selected in this paper refers to the recorded population with huge wealth. These people basically from the Forbes Fortune list, the Hurun Fortune list and the Chinese Fortune 500 list, etc. E.g. Ma Yun, Wang Jianlin and Warren Buffett. Most of them created autobiographies, received media interviews, gave speeches, and published a lot of comments on social media like Sina Weibo. Therefore, data on extremely wealthy populations can be collected in the above aspects.

The text data related to the population with the average economic level above the poverty line, i.e. text data of non-extremely rich populations. In this paper, most of this kind of data comes from Sina Weibo remarks and literary works on major websites. According to the rankings of traditional writers' rich list and online writers' rich income list in China, 2017, except the number one online writer's annual income is 110 million, the rest of the writers' annual income is far less than the annual income of the extremely affluent population. The accumulation of personal wealth of these writers is far less than the wealth of the extremely affluent population. Therefore, the economic status of ordinary writers who are not on the list can be divided into the ordinary economic level. From the above two aspects, the network text data released by the general economic level population can be collected.

The screening criteria after data collection is that the collected texts were created by the creator himself. And we can extract at least 1000 keywords from the text content.

In our research, we collected and selected 64 autobiographies, 6 transcripts of Interviews and speeches, and 110 Public remarks on Sina Weibo of extremely affluent populations, a total of 180. And 77 Public remarks on the Sina Weibo and 239 Internet novels of the general economic level population are collected and selected, a total of 316. We finally collected 496 network text data published by individuals at different economic status.

In our study, ten classified items of the THU Open Chinese Lexicon were selected as features at first. These features included animal, finance, automobile, idiom, place name, food, information technology, law, historical celebrity and medicine. Table 1 shows the contribution rate of the ten features in descending order and the cumulative contribution rate of them.

Table 1 feature's contribution rate and cumulative contribution rate

feature	contribution rate	cumulative contribution rate
animal	0.22717	0.22717
Place name	0.12778	0.35495
food	0.11954	0.47449
information technology	0.10824	0.58273
law	0.10249	0.68522
historical celebrity	0.08611	0.77133
medical	0.08421	0.85554
idiom	0.06763	0.92317
car	0.05433	0.97750
finance	0.02250	1.0

According to the results of PCA, we chose eight features with a cumulative contribution rate of 92.32% to complete data dimension reduction. They were animal, place name, food, information technology, law, historical celebrity, medicine and idiom. According to the PCA, the feature of selecting the cumulative contribution rate of more than 85% can include the principle of the main information in the data, and the above feature is a variable that can reflect most of the information of the data.

Data preprocessing

For the obtained text data, it is first necessary to perform preprocessing such as word segmentation, keyword extraction, keyword classification and keyword frequency statistics. The keyword frequency statistics of each text is a vectorized object, and such object will be used for subsequent training.

1) Word segmentation: Using the jieba library in Python, Chinese word segmentation is performed for each text data.
2) Keyword extraction: Keyword extraction is performed according to the word frequency-inverse document frequency (TF-IDF) of the vocabulary in each text data object after word segmentation. The larger the TF-IDF value of a word, the greater the importance of the word in the text. We extracted the first 1000 keywords of each text data in the experiment.
3) Keyword classification and keyword frequency statistics: In this paper, the 1000 keywords were classified based on the ten types of vocabulary in THU Open Chinese Lexicon. Furthermore, the number of times these keywords appeared in each vocabulary category were counted and saved in an form. If it is a text related to an extremely affluent population, mark 1 after the corresponding object of the word frequency statistics table. If it is a general economic level population related text, mark 0 after the corresponding object of the word frequency statistics table.

Finally, a marked word frequency statistics table is obtained. It is a statistical table of the frequency of occurrence of keywords of each text in 10 vocabulary categories. This statistical table is the result of data preprocessing and is the input data of the subsequent machine learning models.

Result of model train

DOI https://doi.org/2024-JICS-22431 | Generated on 2025-04-05 21:21:36

In the PCA calculations in 3.2., it has been reflected that there is a correlation between the features and the individual economic level, but the ability to predict the individual economic status remains to be verified. The details of the impact of selected features on individual economic levels are also unknown.

According to the above experimental requirements and the characteristics of various machine learning models, we chose logistic regression model. Also we used K-fold cross-validation in the experiment. The evaluation criteria was MMSE.

In the logistic regression model, the parameters to be adjusted are the number of experimental sample groups in the k-fold cross-validation and the coefficient of the regular penalty. Assuming that the number of experimental sample groups in the k-fold cross-validation is CV, the coefficient of the regular penalty is C. In the experiment, we used undetermined coefficient method to adjust the parameters. First, the CV value was fixed, and the C value was changed to obtain the model's predictive accuracy rate. The optimal coefficient C appeared at the highest accuracy rate of the model. Next, the C value was fixed, and the optimal CV value corresponding to the highest accuracy was obtained by adjusting the CV value. Thereby we got the optimal parameter combination.

Figure 2 is an image of accuracy rate obtained by adjusting the CV parameters in the k-fold cross-validation when the regular penalty term's coefficient is 3. The abscissa is the fold in the k-fold verification, and the ordinate is the model accuracy under the corresponding fold. It indicates that in the case where the coefficient of regular penalty is 3, when the CV value is between 65 and 70, the model accuracy is the highest, about 75%. Thus we took the CV value of 69.

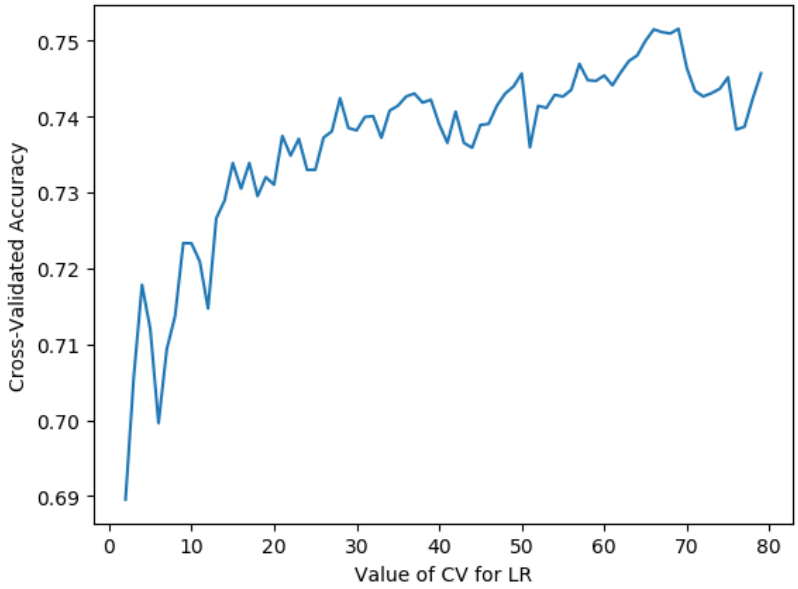

Fig. 2: Accuracy with CV value changing

Figure 3 is an image of predictive accuracy rate obtained by adjusting the C value in the case where the CV value is fixed to 69. The ordinate is the model prediction accuracy, and the abscissa is the coefficient of the regular penalty term. It shows that the accuracy of the logistic regression model has stabilized in the above situation, and it is about 75%. We took the C value of 1.

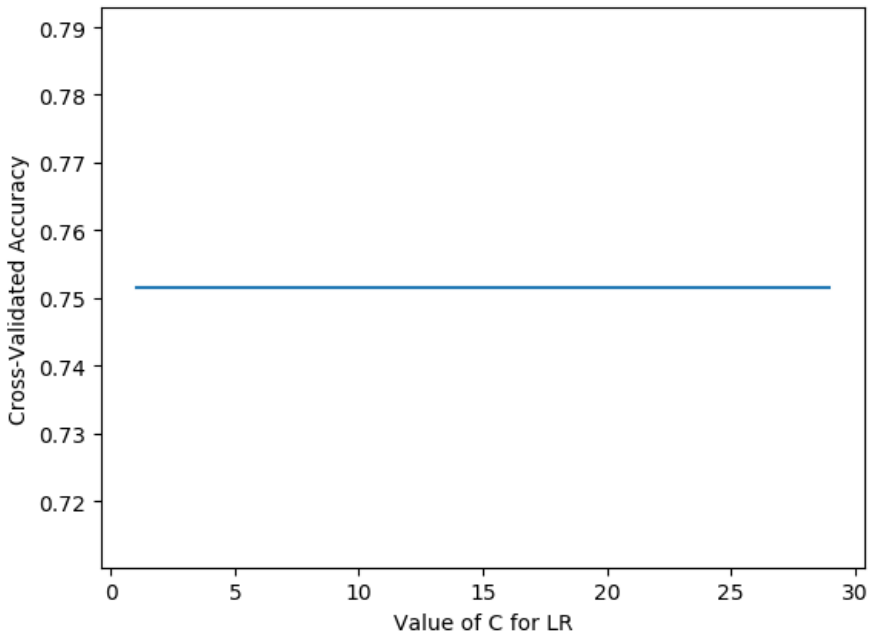

Fig. 3: Accuracy with C value changing

As can be seen from the above figure, the accuracy of the model has not changed with the change of the regular penalty term coefficient. It can be inferred that the model has been well fitted at a k-factor coefficient of 69, and the weight of each input feature data has been stable. That is, the influence of each feature on the experimental results has been calculated, and the addition of the regular term has not affected the experimental results too much.

Therefore, it can be concluded that in the case of the optimal coefficient combination, that is, when CV is 69, C is 1, the minimum training mean square error of the model is 24.48%, and the minimum mean square error of the test is 24.84%. The prediction of the model is good. Table 2 shows the error rate of training and testing when the CV value is 69 and C value is 1.

Table 2 Training error and Test error

Training error	Test error
24.48%	24.84%

At the same time, we also obtained the weight of the model input variables under the minimum mean square error, and the weight of each feature represents its influence on the model output. The larger the absolute value of the weight, the greater the influence of the variable on the output. If the weight is a complex number, it has a positive impact on the prediction result. If the weight is a complex number, it has a negative impact on the prediction result.

The resulting model is:

$$y = \frac{1}{1+e^{-f(x)}} \tag{4}$$

$$\begin{aligned} f(x) = &-0.01473988x_1 + 0.36849211x_2 + 0.45902708x_3 \\ &+0.13163442x_4 + 0.49192715x_5 + 0.43020734x_6 - \\ &0.04042599x_7 - 0.31837514x_8 - 0.99467444 \end{aligned} \tag{5}$$

The above result shows that among the eight selected features, five features such as idiom, place name, food, information technology and law are positively correlated with the individual economic level. Among them, vocabulary of information technology has the greatest impact on individual economic level predictive results, followed by place names and law. Features include medicine, historical celebrities and animal are negatively correlated with the experimental results, with vocabulary of medicine having the greatest impact on individual economic levels.

4. Conclusion

The experimental results show that the internet text data such as autobiographies, remarks on Sina Weibo, transcripts of interviews and speeches, literary works, etc. can reflect a person's economic situation to a certain extent. The vocabulary of information technology has the greatest positive impact on predictive results, followed by place names and law. The vocabulary of medicine has the greatest negative impact on individual economic levels. And the individual economic level is effectively predicted though network text data with an accuracy rate of 75.16%. Compared with the traditional ways of using statistical methods to obtain the economic level of groups, the method of fitting and predicting the network text data by using the machine learning models not only has the advantages of high speed, short cycle, real-time, saving economic costs etc., but also provides a convenient and fast new way to collect groups' economic status.

5. Prospect

The data used in the study is web text data published by population of different economic levels. However, during the experiment, only the data of extremely affluent population and the general economic population were collected, the data of the extremely poor population cannot be collected quickly and easily. Therefore, it is temporarily impossible to analyze the data of the extremely poor. Another problem is that only two types of economic grading are considered in the study. In the future experimental exploration, the economic grading can be appropriately increased or the grading threshold can be calculated to achieve the multi-division of populations with different economic levels. These outlooks are all issues that need to be considered in the future to explore such topics.

6. Acknowledgements

This research is supported by MOE(Ministry of Education in China)Project of Humanities and Social Sciences(No. 17YJC190035), Science and Technology Department of Sichuan Province, Fund of Science and Technology Planning (No. 2018JY0290), Meteorological Information and Signal Processing Key Laboratory of Sichuan Higher Education Institutes(NO. QXXCSYS201606). And we use Tsinghua Open Chinese Lexicon in our research. We are very grateful to the above institutions for their support of this experiment.

7. Reference

[1] Choi, Hyunyoung, *Predicting Initial Claims for Unemployment Benefits*, Available at SSRN: https://ssrn.com/abstract=1659307, July 22, 2009.

[2] Kuethe, Todd H. & Hubbs, Todd & Sanders, Dwight R., *Evaluating the USDA's Net Farm Income Forecast*, Journal of Agricultural and Resource Economics, Western Agricultural Economics Association, vol. 43(3), September, 2018.

[3] David E. Bloom, David Canning, Günther Fink, Jocelyn E. Finlay, *Does age structure forecast economic growth?*, International Journal of Forecasting, Volume 23, Issue 4,2007,Pages 569-585,ISSN 0169-2070.

[4] Hsiao-Tien Pao, *Forecast of electricity consumption and economic growth in Taiwan by state spacemodeling*, Energy, Volume34, Issue11, 2009, Pages 1779-1791, ISSN0360-5442.

[5] Su Z., *Chinese Online Unemployment — Related Searches and Macroeconomic Indicators* [J], Frontier Economics of China, 2014, 9(2) : 347 — 376.

[6] Letouzé E, *Big Data for Development: Challenges & Opportunities* [R], New York,USA: UN Global Pulse,2012.

[7] Shiyi Han, Yuhui Zhang, Yunshan Ma, Cunchao Tu, Zhipeng Guo, Zhiyuan Liu, Maosong Sun. *THUOCL: Tsinghua Open Chinese Lexicon. 2016.*