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