HOLISTICA Vol 8, Issue 3, 2017, pp. 55-60
DOI:10.1515/hjbpa-2017-0023

Modelling prediction of unemployment statistics using web technologies

Mioara, POPESCU,
Bucharest University of Economic Studies, 6, Piața Romană, district 1
mio.popescu@yahoo.com

Abstract

The global diffusion of Internet involves economic, political and demographic factors that can predict in real time. In this article, we demonstrate that according to data provided by EUROSTAT, the number of people looking for a job in Romania it is correlated with specific query terms using Google Trends. Search engine data is used to "predict the present" values of different economic indicators. The obtained results are compared with the classical method of developing the economic indicators, with official EUROSTAT employment data. In this paper, we demonstrate that the new methods to extract the economic indicators from web technologies are accurate.

Keywords: Data Mining, Big Data, Demography, Unemployment, Job Search

JEL Classification: C53, J11, E24, M31

1. Introduction

There are many governmental agencies and organization that periodically publish various economic indicator in different domains. There is a significant delay of months and years between the data collection phase and the public release of the final indicators.

In the last decade, we observed a need for real time tools that can make available economic data. Google, Apple, Facebook, Amazon and many others use the real-time data generated by the customers' activities to extract knowledge from raw data. In this paper, Google Trends was the tool used as a data source for the real-time index of queries performed by the users on Google search engine. We hypothesized that the queries could be used together with different economic indicators to predict and forecast various economic trends.

The research performed in the last years emphasized that the search engine queries could be utilized as important economic indicators for purchases prediction, even from the phase when the consumers start planning their shopping list. In the same time, a study performed by Jennifer L. Castle et al., 2009 reveals that prediction in itself is a critical operation which also inevitably comes with a series of advanced econometric research questions which are considered as one of the most complex topics in this domain.

2. Literature review

”One of the first approach of Ettredge, 2005 demonstrate that the data mining technologies applied to web data can be successfully used in the prediction of the unemployment indicators.” (Ettredge et al., 2005)

”Also, Cooper, 2005 explored the use of internet search queries in different medical domains and topics. In the following years other studies examined the data generated by the web activities in various fields.” (Cooper et al., 2005)

”A few years later, a study performed by Choi and Varian, 2009 suggested that Google Search Insights can be used as a tool to estimate different economic indicators.” (Choi and Varian, 2009)

”Recently, Baker and Fradkin, 2011 have used the same tool to examine the unemployment indicator.” (Baker and Fradkin, 2011)

”In the same year McLaren and Shanbhoge, 2011, observed that the web search data can successfully be used by banks to predict different economic indicators.” (McLaren and Shanbhoge, 2011)

3. Google Trends

Another tool which provide time series of queries performed using the search engine in a specific geographic area is Google Trends. The information collected by this tool is provided using a sampling process, with different daily percentages. Due to the extensive privacy restrictions, in general are tracked only specific searches with which have an important volume. The query index provided is available at the country level but also at more regional levels and districts, for several countries.

4. Examples

The current study found that the Google trends time series is a good tool to improve the prediction of unemployment statistics.

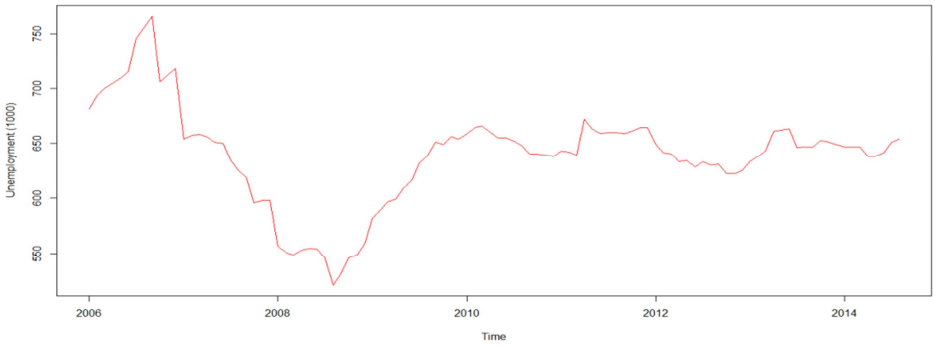
"Data mining technologies are increasingly being used to improve the forecasting in a variety of domains, their excellent results very well known, raised the issue of the robustness and accuracy of those methods" (Lazer, 2014).

The models presented in this paper analyse the unemployment in Romania from two sides one being a Romanian native speaker, it was considerably easier to identify the best query terms used for the job search and the most popular only platforms for job seeking and the other one, there has been changes in evolution trends since 2009 till nowadays. Unemployment has been dropping since 2009 mainly due to the emigration of unemployed persons and also a considerable number of people that stopped looking for a job.

4.1. Available data

"Official data are displayed monthly number of unemployed people, non-seasonal adjusted, from January 2006 to August 2014" (source: Eurostat).

Figure 1. Unemployment Rate (thousands)

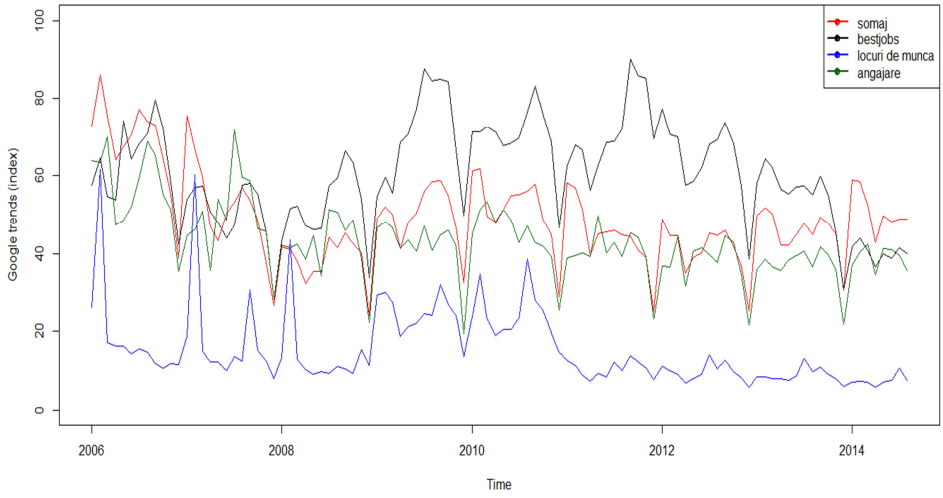


Source: EUROSTAT

Indexed data for the most four popular terms regarding unemployment: (somaj), (bestjobs), (locuri de muncă), and (angajare).

As shown in Figure 1, a fall in the number of unemployed persons was in Romania since end 2006 until mid-2008, there was an increase until 2010 and somehow stable from 2011 until mid-2014.

Figure 2. Google trends (index); January 2006 – August 2014



Source: Google trends

4.2. Models

The analysed model is based on an autoregressive (AR) model. The model is applied to forecast the unemployment in a specific month (t), using the data from the previous month (t-1).

$$\log(y_t) = a + b\log(y_{t-1}) + e_t$$

where y_t is the unemployment rate at month t, a and b are the variables to estimate, and e_t is the error coefficient.

Another model is the baseline adjusted on the selected query terms:

$$\log(y_t) = a + b_1\log(y_{t-1}) + b_2(somaj)_t + b_3(bestjobs)_t + b_4(locuri\ de\ munca)_t + b_5(angajare)_t + e_t$$

where a and b_i are coefficients and $(q)_t$ is the search volume of the query term q.

The four query terms which people search on Google when unemployed in Romania:

- “somaj”, “locuri de munca” and “angajare” are the most used queries when unemployed people perform internet searches to find resources for improving their living condition;
- “bestjobs” is the most popular website for job hunting in Romania;

The data were downloaded on 30 September 2014 and the R software was utilized to process and analyse the data.

5. Results

As we can observe from Figure 1 the time series predicted by the models together with the official data; models were adjusted using the entire dataset (from January 2006 to August 2014).

Also, the adjustment model shows better the prediction than the AR model. Then, for each month after August 2011 two models were adapted in all previous months (i.e. from December 2008 to $t-1$) and the forecast is carried out at month t.

For each month after August 2011 the data was adjusted in two models for the previous months (from December 2008) and the prediction was implemented at month t.

From Figure 2 can be observed that the adjusted model obtained with data provided by Google Trends has better results which fits the official data provided much better than the simple model.

6. Conclusions

The results show using the data from the search engine like Google it is feasible to forecast the unemployment rate faster close to the traditional method of data collection and publishing by the governmental agencies and public institutions. Also, we can observe that by adjusting an AR model it is probable to improve the forecasting accuracy.

The baseline model used in this study is very simple and detailed investigations with other models are required to have improved outcome.

Even if this model is not perfect, the results obtained can be used by the public and private institutions as an indicator that can give an idea of what should be expected in the short term regarding the evolution of unemployment levels.

References

[1] Castle, L., Fawcett, W. P. and Hendry, F. (2009). Nowcasting is not just contemporaneous forecasting. National Institute Economic Review, pp. 71-89, URL http://ner.sagepub.com/content/210/1/71.abstract.

[2] Cooper, C., Mallon, K., Leadbetter, S., Pollack, L. and Peipins, L. Cancer internet search activity on a major search engine, United States 2001-2003. J Med Internet Res, 7, 2005. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1550657/.

[3] Choi, H. and Varian, H. (2009). Predicting the present with Google Trends. Technical report, Google. [Online] (URL http://google.com/googleblogs/pdfs/google_ predicting_the_present.pdf). (Accessed 30 September 2014).

[4] Choi, H. and Varian, H. (2011). Predicting the present with Google Trends. [Online] (URL http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf). (Accessed 30 September 2014).

[5] Ettredge, M., Gerdes, J. and Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. Communications of the ACM, pp. 87-92. URL http://portal.acm.org/citation.cfm?id=1096010.

[6] Holmes, E.E., Ward, E.J. and Scheuerell, M.D. (2014). Analysis of multivariate time-series using the MARSS package. [Online] (URL http://cran.r-project.org/web/packages/MARSS/vignettes/UserGuide.pdf). (Accessed 30 September 2014).

[7] Lazer, D. M., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: traps in big data analysis.

[8] McIver, D.J., Brownstein, J.S., (2014). Wikipedia Usage Estimates Prevalence of Influenza- Like Illness in the United States in Near Real-Time. PLoS Comput Biol 10(4): e1003581

[9] Nick McLaren and Rachana Shanbhoge (2011). Using internet search data as economic indicators. Bank of England Quarterly Bulletin. [Online] (URL http://www.bankofengland. co.uk/publications/quarterlybulletin/qb110206.pdf). (Accessed 30 September 2014).

[10] Perduca, V., Ferreira P., (2014). Improving prediction of unemployment statistics with Google trends.

Scott Baker and Andry Fradkin. (2011) What drives job search? Evidence from Google search data. Technical report, [Online] (URL http://papers.ssrn.com/ sol3/papers.cfm?abstract_id=1811247). (Accessed 30 September 2014).