A PUBLIC ISSUE PROBING SYSTEM BASED ON SEARCH VOLUME INDEX WITH ATTENUATION OF MEDIA COVERAGE EFFECT

Shueh-Cheng Hu
Department of Computer Science & Comm.
Engineering Providence University
Taichung 43301, Taiwan, ROC

I-Ching Chen
Department of Information Management
Chung Chou University of Science and Technology
Chang Hua 51003, Taiwan, ROC

Abstract — When people widely use online search services to find out information of concern or interest, the search volume index (SVI) can serve as an indicator for nowcasting or even forecasting some phenomenon and issues, such as epidemic diseases, abnormal economic trend, consumers’ behaviors, etc. However, if the search volume index is significantly influenced by media coverage, its appropriateness of nowcasting problems that really concern the public, will be questioned. In light of most people still rely on search service to find out information of concern, as well as media coverage does influence search volume index, the present work aims to research and develop a Web-based probing system, which can identify problems concerning the public mostly with attenuation of media coverage effect. With it, administrators of public affairs can gain insight into noteworthy search trends and the issues behind the trend, which will facilitate their corresponding decision making.

Keywords – Public issues, search volume index, media coverage, regression analysis, R language.

1. INTRODUCTION

To respond public affairs efficiently and effectively, administrators usually ask themselves a question: “what are the problems concerning people mostly?” Traditionally, public affairs administrators are able to understand issues concerning most citizens by using questionnaires, phone interviews, conferences, etc. In the age of Internet, many people rely on search engines to find out information regarding their interests and concerns, correspondingly, issues concerning public could be to certain extent reflected by collective searching behaviors. The collective searching behaviors could be measured by the so called search volume index (SVI) [1], which could be retrieved from online search service providers such as Google. Several prior works reported that SVI had been successfully used to nowcast or forecast the spreading of diseases [2], trend of financial market [3, 4], and sales of products and services [5]. The fundamental assumption of the SVI being able to precisely nowcast or forecast a particular issue is that most people conduct Web search due to their self-awareness or self-interest regarding the issue, rather than pure curiosity. As Ginsberg et.al stated in their work detecting influenza epidemics via Web query
they assumed that people who conduct Web search by using specific keywords do have motivation to avoid or cure diseases, rather than just for curiosity or fun. However, many people start Web searching only after receiving relevant messages from media, and in this case, people conduct Web search because of not only their self-awareness and interests, but also curiosity, thirst for knowledge, as Figure 1 illustrates. In consequence, the SVI cannot constantly serve as a precise measure to nowcast and forecast issues that really reflect public concerns.

Figure 1. Factors influencing public Web search

In light of media coverage’s influence on search volume index that can reflect emerging issues concern public, the present work aims to propose a Web-based system for probing public issues, but with attenuation of media coverage effect. Besides, the corresponding design and implementation are presented. The major benefit of the system is that it can help public affairs administrators in identifying the most significant public issues in time, thus can prepare, and handle them promptly.

II. PRIOR WORKS REVIEW

Comparing with traditional methods to understand and measure public opinions or concerns such as telephone survey, people quickly found that SVI is a good option for nowcasting or forecasting social phenomena and public issues, because of its many merits including immediacy, easy availability, standardized numerical format (fully automatic processing), etc.

Herman Anthony Carneiro and Eleftherios Mylonakis built a system for monitoring disease outbreaks based on the Google trends service, and they reported that their system can raise alarms 7 to 10 days earlier than the conventional disease alarming system [1]. Besides public health issues, researchers also found that the SVI also serve as a good indicator for a company’s stock price and trade volume while the economy condition is stable, especially for technology companies [7]. Researchers also proved that the SVI outperform other survey-based indicators in forecasting consumers’ behaviors [8].

Brian Weeks and Brian Southwell study the relationships between media coverage and public Web searching behavior, they pointed that the media coverage volume on a particular issue can precisely predict the SVI associated with the same issue, although the media influence is transient [9]. Matthew Ragas, et al. also found that the volume of media coverage on an issue has significant impact on people’s Web searching behavior [10]. Given these facts, the SVI must be adjusted based on the influence brought by media coverage, before it is applied to precisely probe things.

III. A SYSTEM PROBING PUBLIC ISSUES

To probe issues that really reflect public concerns instead of curiosity or other factors influencing search volume index, a system as shown in figure 2 was proposed accordingly. There are the modules for retrieving raw data from search service provider, modules for alarming users and presentation, and the most important one, a module for modulating the influence of media coverage based on the results of quantitatively analyzing the relationship between public news acquisition and their Web searching behaviors.
Regarding the implementation works, the R language and a number of its packages were selected due to its broad offerings in multiple facets that include retrieve SVI data from Google trends (offered by the package: gtrendsR), regression analysis (offered by packages lm, glm, etc.) time series analysis (ts, zoo, tseries, etc.), Web-based software architecture (shiny), and data visualization (ggplot2 and ggvis), which are all essential to functionality of the present system.

IV. MODULATION WORK BASED ON QUANTITATIVE ANALYSIS

To adjust the influence of public news acquisition on their Web searching behavior, it is necessary to quantitatively measure the relationships between public news acquisition and their Web searching behaviors. For nowcasting purposes, regression analysis could be applied, while the time series packages can address the forecasting purposes.

Taking a nowcasting case as an example, to investigate the relationships between the news acquisition and Web searching behaviors on the term: "cancer" in Taiwan during the last 5 years, 3 different regression (linear, polynomial, power) models were tried, based on the observation of regression line and the scatter plot of subject data, linear model is a better option for describing the relationship between the explanatory (Cancer_News) and response variables (Cancer_Web). By applying the lm package in R, we can easily obtain a regression equation that describes the linear relationship between the two variables:

\[ \text{Cancer}_\text{Web} = 38.09487 + 0.23609 \times \text{Cancer}_\text{News} \]

Before applying the regression model to adjust the influence of news acquisition, it is necessary to check the quality and fitness of the regression model according to the output provided by the R packages. The F-test for overall significance in the analysis of variance (ANOVA) indicates that whether we can reject the null-hypothesis and conclude that the regression model provides a better fit than the intercept-only model. The significance of a predictor variable could be judged according to its p-values. The corresponding coefficient represents that the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.

Linear regression using ordinary least squares identifies a formula that minimizes the distances (residuals) between the fitted line and all of the data items. R-squared, also known as the coefficient of determination, is an instrument for measuring how close the data points are to the fitted regression line, it is the percentage of the response variable variation that could be explained by the regression model. The adjusted R square can show whether an adequate value exist for predicting human behaviors or social phenomenon, which usually are not predictable as science phenomenon.

To further check the fitness of the model, a residual plot of a regression model is useful because that plot can shows whether residuals was either systematically high or low, i.e., if there was randomness in the residuals. R-squared, also known as the coefficient of determination, is an instrument for measuring how close the data points are to the fitted regression line, it
is the percentage of the response variable variation that could be explained by the regression model.

**V. CONCLUSIONS**

A. Contribution and Anticipative Impact

The search volume index, which is available through search service provider, has been successfully used to nowcast and forecast some social phenomenon or human behaviors. However, because media coverage effect, it is necessary to modulate the influence of public news acquisition on their Web searching behavior that is measured by the SVI. To address this issue, the present work proposes a system that can probe pubic issues real-time while taking the effect of media coverage into account. Using this kind of system, public affairs administrators can efficiently and effectively grasp issues that really concern the public in depth, prompt and dead on target resolutions thus could be proposed accordingly.

B. Future Directions

The prior findings pointed that the major influential factor of public Web searching behavior is media coverage, which justifies the application of simple regression technique while building a model for nowcasting things based on the SVI. However, it is worthy to explore whether there other factors that lead to the fluctuation of collective Web searching behaviors, and the corresponding statistical technique such as multiple regression analysis should be used accordingly.

The regression analysis results for queries in different subject domains vary, in other words, there is no common regression formula fitting all conditions. In consequence, public affairs administrators who want to identify and measure public issues by applying the SVI need to conduct regression analysis based on the particular queries of interest. Therefore, it is worthy to develop an automatic and adaptable procedure for conducting volume of regression works that are relevant to public issues.

Moreover, to find out whether there is any causal relationships between the media coverage and public Web search behavior, advanced statistical techniques such as time series analysis could be applied.

**ACKNOWLEDGMENT**

This research work has being funded by the grant from the Ministry of Science and Technology, Taiwan, ROC, under Grant No. MOST 105-2221-E-126-012-. We deeply appreciate their financial support and encouragement.

**REFERENCES**


