Public Opinion between Blogsphere and Real World

Heng Lu
Web Mining Lab, Dept. of Media & Communication, City University of Hong Kong

Abstract

Blogs are arguably the work horse of social media as the opinions expressed there are more articulated and elaborated than other platforms (e.g., microblogs or online forums). However, it has remained unknown the extent to which blogs represent views of the general public. In the paper, I compare the views as expressed in randomly sampled blogs and gauged by real world statistics. Using the number of tourists for provinces and Chinese Music Chart billboard as benchmarks of real world, I assess the correlation between blog-searching results and real world statistics. The results indicate that it is vulnerable to use blog-searching results to project the real world statistics, especially for those statistics without dramatic fluctuations.

Keywords:

Blog mining, blog searching, public opinion, social media

Introduction

Although survey is the most promising practice to accomplish accurate representative public opinion, it bears several disadvantages. Since 1990, the response rates continue to deteriorate; the costs of face-to-face surveys continue to inflate; the coverage rate of traditional telephone survey frames continues to decline (Groves, 2011). The rise of User-Generated Content (UGC) offers a great opportunity for researchers to detect, assess, and predict online public opinion among many other Xs; for example: movie box (Asur & Huberman, 2010), stock market (Bollen, Mao, & Zeng, 2011), and electoral votes (Tumasjan, Sprenger, Sandner, & Welpe, 2011). Blogs are arguably the work horse of social media as the opinions expressed there are more articulated and elaborated than other platforms (e.g., Facebook or Twitter). "Blog searching" or "Blog mining" is a rising method of public opinion research (Agarwal & Liu, 2008; Attardi & Simi, 2006; O'Leary, 2011; Thelwall, 2007).

There are several advantages of using blog data for public opinion and other research. Firstly, the blog data is near zero cost. Unlike the expensive costs of household and telephone survey, the data-retrieval is zero cost. A computer with Internet access is adequate for crawling blog data.

Secondly, the blog data contains time stamps, and therefore, is desirable for longitudinal studies. Longitudinal survey data is precious due to the inflating costs and declining response rates, especially for panel surveys with many waves. The blog data with timestamps record a series of individual behaviors. Researchers could trace the trajectory of individual behaviors over a desirable time span.

Thirdly, the blog data could be very large in volume. A well designed sampling strategy and a fine crawler could easily get records of user activities and contents of millions of blog users. Some blog service providers like twitter even provide their Application Programming Interfaces (APIs), which enables researchers to obtain well organized data.

Fourthly, the opinions expressed on blogs are usually more elaborated than that on other online UGC platforms (e.g., Facebook or Twitter). The first three advantages mentioned above are commonly shared by all online UGC data sources. Only the fourth advantage is unique for blog data. The blog posts contain more detailed information about the opinion that could be used for more sophisticated analysis, for example, different dimensions of opinion.

There are also several disadvantages of using blog data for public opinion research. Firstly, drawing a representative sample is difficult. Currently, most of the work is based on non-probability samples. Although the sample size is extremely large (i.e.,

Copyright is held by the authors.
The annual conference of the World Association for Public Opinion Research, Hong Kong, June 14-16, 2012.
Correspondence should be addressed luheng9@gmail.com.

usually in millions), there is no warranty for its representativeness.

Secondly, user demographics of blog websites are not necessarily match that of neither online population nor offline population. It is fine to use blog data source to study the public opinion of bloggers. It should be caution, however, to infer the results to all Internet users, especially to the real world population.

Thirdly, self-selection bias lies behind the blog data. People publish their blogs voluntarily; therefore, the blog data is produced by people of certain same characteristics (e.g., sharing experiences and opinions actively).

Related Work

The main tasks of "blog mining" are: "retrieving opinionated postings and their polarity, and retrieving topical items both at posting and feed level" (Demartini, Siersdorfer, Chelaru, & Nejdl, 2011, p. 466). Researchers from the field of Information System and Computer Science have done a lot of pioneer work following their research tradition of information retrieval and text mining.

People write blogs to document their daily lives, to comment on events, to discuss their interests. Mining blogs might therefore reveal information about the following three questions: "What can we learn from blogs?" "What is the 'sentiment' (positive or negative) about certain issue?" and "Can these blogs, individually or in the aggregate, monitor or forecast the public opinion/image of certain events or entities (e.g., products and politicians) over a considerable timespan?"

The exemplary work of answering the question of "what can we learn from blogs" is done by Gordon and his colleagues. They try to find "commonsense knowledge" or "commonsense causal reasoning" by analyzing the blog posts. They point out phrasing associated with causal relationships (such as "I did X so then Y happened") from millions of blog posts (Gordon, 2010).

The latter two questions are more relevant to public opinion research. A lot of work has been done to answer these two questions. For example, O'Connor and his colleagues find that the sentiment word frequencies in 1 billion Twitter messages could reflect the trajectories of consumer confidence measured from polls over 2008 to 2009. However, concerning the presidential candidate preferences, they report a low correlation between the aggregated sentiment over those tweets and results from election polls (O'Connor, Balasubramanyan, Routledge, & Smith, 2010). Han Woo Park collects blog posts related to 29 candidates for the 2009 Korean National Assembly by-election. The results of correlation analysis (Spearman rho=.797, $p<.01$, N=29) significantly indicate the positive relationship between blog postings and vote. The results of a simple regression analysis indicate that the number of blogs by candidates can be regarded as a significant determinant of the number of votes (Park, 2010). Thelwall (2007) proposes to use blog search engines as new source for social scientists to track debates or for retrospective public opinion information. In the case (i.e., Danish Cartoons) he demonstrates, he finds the simplest techniques of content analysis is more effective at getting useful information than the more sophisticated time series scanning.

Most of the previous studies are single event/case-driven. They usually focus on one single event (e.g., 2008 presidential election) or single entity (e.g., Walmart). They monitor and forecast people's attitudes toward the research target over certain time span. Usually there are dramatic fluctuations of public attitudes towards these entities and events; and therefore, they come into the scope of these studies. However, very few studies use blog data to project the relative stable real world statistics.

In this study, I use the data of tourism and music charts as real world benchmarks. People talk about their travel experiences and favorite artists in their blogs. Usually, there would not be dramatic fluctuations in these statistics.

Case 1: Blog Searching and Tourists

Data

The real world data is *the number of tourists* for each province in mainland China over the year 2006 to 2009. These statistical numbers are provided by the National Bureau of Statistics of China. The *rank of tourists* is the rank of *the number of tourists* for each province over the year 2006 to 2009.

Over 200 thousand blog articles of 10 thousand bloggers of Sina Blog constitute the online data. Sina Blog (http://blog.sina.com.cn) is the biggest blog service provider in China. These 10 thousand bloggers are randomly selected using Random Digit Dialing method (Zhu, Mo, Wang, & Lu, 2011). All these bloggers post at least two blog articles. The publish date of these blog articles ranges from January 1, 2006 to December 31, 2009. I search the names of all provinces/municipalities and all prefectural-level cities in these blog articles. The numbers of occurrences of each above province/city names in these blog articles are recorded. 1.01% of all blog articles contain at least one province/city name. Then I aggregate these numbers to the provincial level. The *frequency of embedded occurrences* for province *i* is the number of blog articles which mention the name of province *i* or the name(s) of prefectural-level cities in province *i* over year 2006 to 2009. The r*ank of tourists* is the rank of *the number of tourists* for each province; and the *rank of embedded occurrences* is the rank of the *frequency of embedded occurrences* for each province. Both *rank of tourists* and *rank of embedded occurrences* are numbers in the range of 1 to 31.

Findings

The correlation between the *number of tourists* and the *frequency of embedded occurrences* is significant (Pearson r=.736, p<.000; Spearman rho=.806, p<.000, N=124); so it is between the *rank of tourists* and the *rank of embedded occurrences* (Pearson r=.827, p<.000; Spearman rho=.827, p<.000, N=124). The blog searching seems to be qualified to be used as a retrospective source of the real world tourist traffic.

I further examine the correlations between the *number of tourists* and *provincial GDP* as well as between the *rank of tourists* and *rank of provincial GDP*. Both correlations are statistically significant too (Pearson r=.879, p<.000 and Pearson r=.890, p<.000; Spearman rho=.897, p<.000 and Spearman rho=.890, p<.000, N=124). According to the correlation coefficients, the GDP is a better predict of tourists.

I conduct simple linear regressions to test the unique contribution of the blog searching results. The semi-partial R square for the *frequency of embedded occurrences* is 0.02 while the R square for the *frequency of embedded occurrences* and *provincial GDP* is 0.79 (See Table 1). The semi-partial R square for blog searching increases to 0.05 when I set the *rank of tourists* as dependent variable and *rank of provincial GDP* and *rank of embedded occurrences* as independent variables (See Table 2).

Table 1 & Table 2 about here

The results of the above extremely simple analysis demonstrate that there is a statistically significant between the results of blog searching and real world statistics such as tourist traffic. The results of blog searching as predictors, however, are not better than some easily found traditional predictors (e.g., GDP in the above case). Furthermore, they have very limited unique contribution to the prediction.

Case 2: Blog Searching and Music Chart

Data

The offline data comes from the Chinese Music Chart (CMC, 中国歌曲排行榜). This is the most popular Chinese music billboard. The Chinese Music Chart releases the most popular Chinese songs and their artists on weekly basis. I focus on the artists of the top 10 songs shown on this music charts from the 37th week in 2005 (the Sina Blog is officially launched this week) to the 52th week in 2009. The CMC releases 164 billboards in these weeks. There are more than 200 unique artists listed on the CMC billboards at least once. The *billboard-score* of artist *A* in week *n* is the square root of her ranking (reverse coded). For example, *Wang Rong, Sun Nan* are the top 1, 2 artists in the 38th week 2005; the *billboard-scores* of artist *Wang Rong* and *Sun Nan* in this week are therefore 3.16 (i.e., sqrt(10)) and 3 (i.e., sqrt(9)).

The online data is same as used in Case 1. The publish date of these blog articles are from the 37th week in 2005 to the 52th week in 2009. I search all the artists ever listed on the CMC billboards in the 200 thousand blog articles. The *occurrence-score* of artist *A* in week *n* is the square root of the number of articles mentioned her in this week. For example, the name of *Zhang Jie* appear on 16 unique blog articles in the 27th week 2007; the *occurrence-scores* of *Zhang Jie* in this week is therefore 4 (i.e., sqrt(16)).

Findings

I test the overall correlation of the online and offline data firstly, and then compare the online and offline trajectories of the same artists from 2005 to 2009.

For each artist, there would be a *billboard-score* and an *occurrence-score*, which arguably reflects the real- world and online popularity of the artist. Putting all the artists together, the correlation between the *billboard-score* and *occurrence-score* is significant (Pearson r=.338, p<.000, N=200). The correlation coefficient is relative small. Nevertheless, the result of correlation test provides face validity for using blogosphere as an alternative source to reflect the popularity of artists.

The time series nature of both data source provides desirable opportunity to test whether or not the online and offline trajectories match each other in a long run. Two artists are in the list of top 10 artists with highest *billboard-score* as well as highest *occurrence-score*. The trajectories of these two artists, Li Yuchun (李宇春) and Wang Lihong (王力宏) are selected to demonstrate the coevolution of artists' online and offline popularity. Li Yuchun and Wang Lihong appear on the CMC billboard top 10 list for 39 and 34 weeks. Figure 1 and 2 show the popularity-trajectories of Li Yuchun and Wang Lihong.

Figure 1 about here

The correlation coefficients between *billboard-score* and *occurrence-score* for Li Yuchun and Wang Lihong are .006 (p=.939, N=164) and -.077 (p=.329, N=164). As we can learn from both Figure 1 and 2, even for the most popular artists as selected in this case, the names of artists are not often mentioned by the 10 thousand bloggers nor listed on the CMC billboard. The *occurrence-score* does not sensitively reflect the trends of *billboard-score*. It is more like a random time series without any trends or cycles.

Conclusion and Discussion

Both cases provide weak, although statistically significant, evidences for the correlation between blog-searching results and real world statistics. From Case 1, we can draw a conclusion like: the more the names of provinces/cities are mentioned by blog articles, the more the tourists are for the provinces. This conclusion, however, is not an important finding because GDP is an easily accessible predictor with larger predictive power. After controlling for GDP, the unique contribution of blog-searching results is quite small.

We can learn from Case 2 that there is a face-validity to use blog-search results to monitor the popularity of artists which is reflected on billboards. However, it is a bad idea to use blog data to scan the popularity-trends of artists. The popularities of artists are reflected by the number of times she is mentioned by blog articles. This reflection is not time sensitive. The frequency of her name mentioned by blog articles could not predict the fluctuations of her popularity reflected on billboards. That may be because the occurrence of artists on billboard is a zero-sum game and on blog posts is not.

The method used in this study is extremely simple. There would be certain noises existing in the results of blog-searching. Some irrelevant blog articles could be filtered out from the sample based on manually coding or more sophisticated text mining technologies. However, the results could still indicate that it is vulnerable to use blog-searching results to project the real world statistics, especially for those statistics without dramatic fluctuations.

Nevertheless, we still face two big challenges to establish blog mining as an alternative way of assessing public opinion. First of all, it is difficult to find a commonly accepted way of interpreting reality. I use the number of tourists and CMC billboard lists as benchmarks of reality. The matching results between blog-searching and real world statistics might be varying across different issues, events, and entities. Secondly, it is extremely difficult to establish a valid and reproducible method of monitoring and forecasting real world statistics by blog mining. We find statistically significant correlations in both cases. We report the results very carefully rather than exaggerate them.

References

Agarwal, N., & Liu, H. (2008). Blogosphere: research issues, tools, and applications. *ACM SIGKDD Explorations Newsletter, 10*(1), 18-31.

Asur, S., & Huberman, B. A. (2010). *Predicting the Future with Social Media.* Paper presented at the

IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

Attardi, G., & Simi, M. (2006). *Blog mining through opinionated words.* Paper presented at the fifteenth Text REtrieval Conference (TREC).

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 1-8.

Demartini, G., Siersdorfer, S., Chelaru, S., & Nejdl, W. (2011). *Analyzing Political Trends in the Blogosphere.* Paper presented at the Fifth International AAAI Conference on Weblogs and Social Media.

Gordon, A. S. (2010). *Mining commonsense knowledge from personal stories in internet weblogs.* Paper presented at the First Workshop on Automated Knowledge Base Construction, Grenoble, France.

Groves, R. M. (2011). Three Eras of Survey Research. *Public Opinion Quarterly, 75*(5), 861-871.

O'Leary, D. E. (2011). Blog mining-review and extensions: "From each according to his opinion". *Decision Support Systems, 51*(4), 821–830.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). *From tweets to polls: Linking text sentiment to public opinion time series.* Paper presented at the Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC.

Park, H. W. (2010). Mapping Social, Political, And Scientific Landscape Using Webometrcs Method. 2012, from http://www.slideshare.net/hanpark/mapping-social-political-and-scientific-landscape-using-webometrcs-city-univ-of-hong-kong-24-march2010

Thelwall, M. (2007). Blog searching: The first general-purpose source of retrospective public opinion in the social sciences? *Online Information Review, 31*(3), 277-289.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election Forecasts With Twitter. *Social Science Computer Review, 29*(4), 402-418.

Zhu, J. J. H., Mo, Q., Wang, F., & Lu, H. (2011). A Random Digit Search (RDS) Method for Sampling of Blogs and Other User-Generated Content. *Social Science Computer Review, 29*(3), 327-339.

Table 1 Unstandardized OLS regression coefficients (standard errors in parentheses) of Number of Tourists (unit: 10k)

	Model 1	Model 2
Constant	2535.99*** (490.48)	2423.92*** (474.26)
Provincial GDP	.78*** (.04)	.65*** (.05)
Frequency of Embedded Occurrences		3.33** (1.04)
R-Square	.77	.79
N	124	124
R Square Change		.02

*** p<.001, ** p<.01, * p<.05

Table 2 Unstandardized OLS regression coefficients (standard errors in parentheses) of Rank of Tourists

	Model 1	Model 2
Constant	1.76*	.47
	(0.76)	(.70)
Provincial GDP	.89***	.62***
	(.04)	(.06)
Frequency of Embedded Occurrences		.35***
		(.06)
R-Square	.79	.84
N	124	124
R Square Change		.05

*** p<.001, ** p<.01, * p<.05

Figure 1 Billboard-score and Occurrence-Score for Li Yuchun over 2006 to 2009

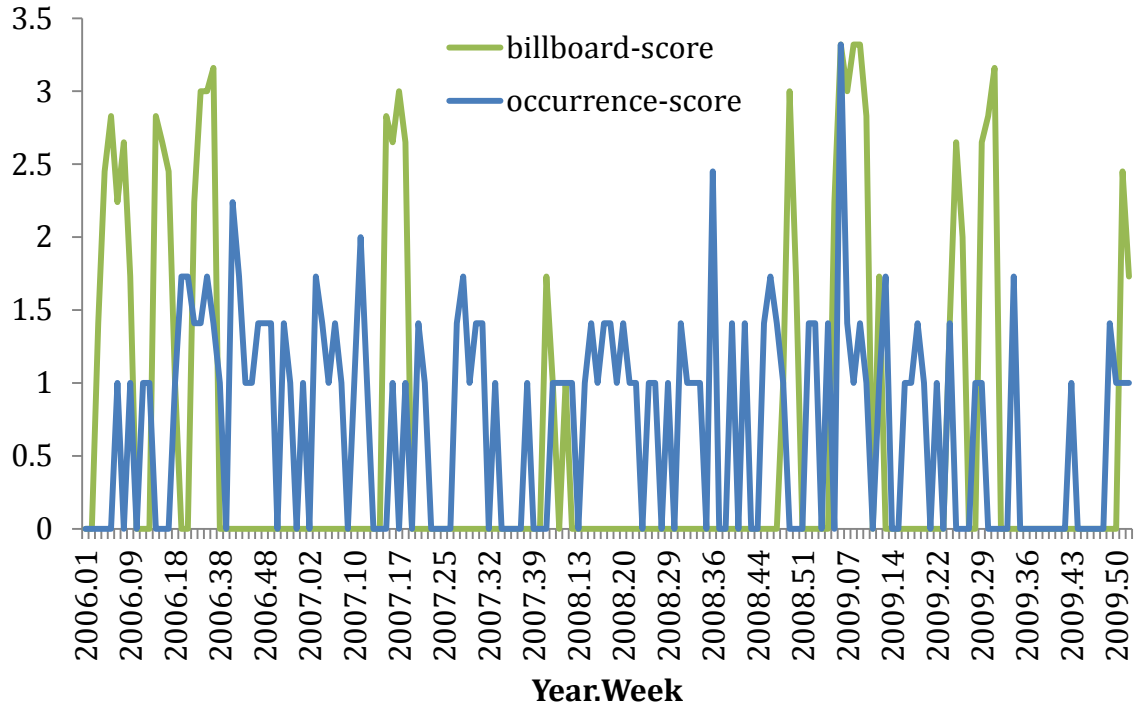

Figure 2 Billboard-score and Occurrence-Score for Wang Lihong over 2006 to 2009