

Yijun Gao
University of Western Ontario, London, ON.

Analyzing News Website from Publicly Accessible Data

Abstract: This study finds some publicly available data, such as the comments posted to the news stories and online survey results, could be an alternative data source for researchers to analyze news websites when the Web server log data are not available.

Résumé : Cette étude indique que les chercheurs pourraient utiliser des données publiques, comme les commentaires de reportages publiés en ligne et les résultats de sondages électroniques, pour analyser les site Web d'information lorsque les journaux transactionnels des serveurs ne sont pas accessibles.

1. Introduction

More and more people rely on the Web to retrieve news and other information. With the help of Web server log analysis techniques, it is easier now for researchers to monitor the general usage and traffic patterns of any news website, and then gauge the online newsreaders' interests, which is of significance to both information scientists and journalists (Thelwall, Vaughan & Bjorneborn, 2005).

In addition to the data generated by servers of the website being investigated, which is almost impossible for outsiders to retrieve, some publicly available data, such as readers' responses to online surveys, Web surfers' comments to news reports, as well as online reprint rates of each individual news items, may offer us some alternative clues in discovering the public's interest in the cyberspace.

According to the *26th Statistical Report on the Internet Development in China*, the number of Web users in this nation exceeded 420 million by the end of June 2010 (CNNIC, 2010). Therefore, China is taking shape as an influential Web power, and should not be ignored by the information science and journalism community.

Due to the commercial sensitivity of the data, few studies have comprehensively analyzed Chinese news websites using Web server logs and survey results, thus leaving some knowledge gaps to be filled.

This paper examined Web server log statistics, online survey results, and top news reports' online reprints rates of the *People's Daily Online* (<http://www.people.com.cn>), a leading comprehensive news portal in China, with the aim of finding alternative method to analyze news website rather than using sensitive server log data.

2. Methodology

A total of 129 days of Web server log statistics of the *People's Daily Online*, including the top 50 Chinese and English news stories with the highest daily pageview numbers,

the number of comments posted to these news items, as well as the top 50 Chinese and English queries submitted to the website search engine of the *People's Daily Online*, were collected from October 2007 to September 2008.

Due to the sheer volume of the raw server log data, the author decided to use the statistical data generated by the WebTrends Log Analyzer installed at the server of the *People's Daily Online*, which efficiently helped us exclude visits generated by “online noises” (e.g. Web crawlers' work, the graphics, banners, menu frames on the webpages, and malicious hits by the hackers or proxies).

Results of an online Chinese survey launched targeting upcoming events in March 2008 were collected after its closing date. Above mentioned Web Server logs statistics from the whole month of March 2008 (31 days) were also retrieved. Online reprint rates for the top 50 Chinese and English news reports were retrieved by Google.com.

All of the top 50 Chinese and English reports/queries, as well as the survey questions collected for this study were classified in accordance with the following scheme: Chinese Politics, World Politics, Business, Sci-Tech and Education, Entertainment, Sports, Health, Culture, Life and Society, Accidents and Disasters, Olympics, Earthquakes, Taiwan, Tibet, Crime, Environment, and Odd.

The relationships between the comment count and the pageview number of each of the top 50 news items were examined, aiming to find whether or not such data will offer us valuable clues to uncover readers' interests. The author also tried to investigate the relationships between survey results and the server log statistics, as well as the ties between pageview numbers and online reprint rates of the top 50 news reports.

3. Results and Conclusions

Frequency distributions of the top 50 Chinese/English news reports' pageview numbers and the top query counts were highly skewed. Therefore, nonparametric statistical tests were used to examine the data sets reported in this study.

The author first tested the frequency distributions of the top 50 Chinese/English news reports measured by the number of stories within each news report's topic category. The same test was also conducted on the top 50 news reports within each of the same topic categories as measured by pageview numbers. Spearman's Rho test revealed remarkably strong positive relationships between the two data sets (all correlation coefficient values were very close to 1).

Thus, it was safe for us to use distribution patterns measured by the number of reports in each category to investigate the general interests of the readers of the *People's Daily Online*. Compared with the pageview number of each news report's category, the number of reports from each category is a publicly available data source for us to access from the leading news websites.

These statistically significant results helped us find an alternative data set for the server log statistics: the most popular news reports list of the *People's Daily Online*, which perfectly reflected the distribution patterns of readers' interests measured by

pageview numbers (almost same ranking).

There were also significant positive correlations ($P < 0.01$, Spearman's Rho test, Correlation Coefficients are 0.692 and 0.757, respectively) between the pageview number of each of the top 50 Chinese/English news reports and the counts of comments the same news report attracted. Thus, it is safe for us to say, at least for the *People's Daily Online*, that comment counts are good alternative data for us to study readers' interests. A significant correlation here also allowed us to make some predictions about the pageview numbers of top news reports based on their comment counts in future studies.

Spearman's Rho test further showed a significant correlation between the "top Chinese news reports' pageview numbers" of the eight matching categories from March 2008 and "survey votes" ($P = 0.037$; Correlation Coefficient = 0.738). Therefore, we can conclude that the survey did reflect readers' interests as shown through the pageview measure. Similar results were also found about the top queries.

There was no significant relationship between the pageview numbers and online reprint rates for each of the top 50 English reports. However, the online reprint rates for each of the top 50 Chinese news items were positively correlated with the latter's pageview numbers. The Correlation Coefficient for the Chinese test was small (0.136), and did not support a strong relationship. Thus, ranking news reports by their online reprint rates did not reflect the readers' interests as measured by the pageview numbers.

4. Discussion and Future Studies

The server logs are too sensitive for the public to retrieve or share, survey results and comment counts would play a somewhat positive role for us to study online readers' interests within the public domain, although there are also some concerns regarding them (e.g. censored survey questions and misleading comments).

However, we could still gauge readers' interests or the public opinion via the comments, because these comments are the words from the real readers and a direct measure of their sentiments. And, leaving a message proves that they have at least browsed some content of the article, rather than mistakenly clicking on it.

Considering all of these limitations, no single method can draw a complete picture of the online news readers' interests. Thus, server logs and surveys should always be employed together in order to make more comprehensive conclusions.

Future research is especially needed to investigate the continuously rapid growth of the "Mobile News Readers," which poses both challenges and opportunities to the press industry in the 21st century. In the meantime, a comprehensive qualitative analysis the readers' comments is also desired to better understand their sentiments both online and offline.

5. References

CNNIC (2010). *The 26th Statistical Report on the Internet Development in China*, by Beijing: China Internet Network Information Center.

Thelwall, M., Vaughan, L. and Bjorneborn, L. (2005), “Webometrics”, in Cronin, B. (Ed.), *Annual Review of Information Science and Technology*, Information Today, Inc., Medford, NJ, 39, 81-135.