Temporal Analysis of the Usage Log of a Research Networking System

Sunmoo Yoon, RN, PhD¹, Sylvia Trembowelski, MS², Richard C Steinman, BA², Suzanne Bakken, RN, PhD¹,²,³, Chunhua Weng, PhD²,³

¹School of Nursing, ²The Irving Institute for Clinical and Translational Research, ³Department of Biomedical Informatics, Columbia University, New York, New York, USA

Abstract

Despite the proliferation of research networking systems (RNS), their value and usage remains unknown. This study aims to characterize the temporal usage of an RNS, Columbia University Scientific Profiles (CUSP), and to inform the designs of general RNSs. We installed a free usage logging service, Google Analytics, on CUSP and applied time series analysis to compare the usage patterns of the two modes of CUSP: restricted (authenticated) and open access. More users searched by person names than by topics, although the latter enables in-depth vertical search of co-author or co-investigator networks for grants or publications. The open-access mode received more page views but less average time spent on each page than the restricted access mode. The numbers of unique users and searches have increased over the time. This study contributes a trend analysis framework for understanding the usage of RNSs and early knowledge of the usage of an open-access RNS.

Introduction

Facilitating interdisciplinary team science is a critical mission of NIH’s Clinical and Translational Science Award (CTSA) program, comprising 60 medical research institutions in 30 states and the District of Columbia in the United States. An important mission of CTSA is to help biomedical researchers identify collaborators. Research Networking Systems (RNS) have been designed in many CTSA institutions to foster collaboration among clinicians and researchers working in multiple disciplines. Understanding information needs of biomedical researchers and collaborator searching methods on RNSs is vital for improving RNSs. One cost-effective way to understand behaviors of biomedical researchers is to analyze web server log files. Log file analysis provides information about system usage, including when, how, where, and by whom the system was used. Although a single method cannot provide a whole picture of user behaviors, previous studies have shown that web usage mining can capture a reasonable amount of information about the performance of a system.

Columbia University Scientific Profiles (CUSP) (http://irvinginstitute.columbia.edu/cusp) is a locally developed RNS that generates a scientific profile for each biomedical researcher affiliated with the Columbia University Medical Center using information from human resources, MEDLINE databases, and university grants databases. We first launched CUSP in March 2011 for internal use by Columbia University employees. In March 2012, we made CUSP open access. During both phases, we used Google Analytics to monitor its usage in real time.

We previously reported the usage of CUSP by authorized users during its restricted access phase. This study aims to characterize and explain the temporal usages of the open-access CUSP through trend analysis and to gain insights for improving CUSP and other open-access RNSs. Specifically, this study addresses three questions: (1) How has CUSP been used? (2) What are the differences in CUSP usage patterns between restricted access and open access modes? and (3) How has usage changed over time during CUSP’s open access period? The Columbia University Medical Center Institutional Review Board approved this study.

Methods

In order to address the first question, we obtained descriptive statistics about CUSP use for the time period from December 2, 2011 to September 19, 2013 from Google Analytics (http://www.google.com/analytics) installed on the CUSP server. The information used for our analysis includes anonymous visitors’ geographical locations, Internet service providers, devices, search terms, the number of page views per visit, visitor status (i.e., new or returning), and the number of visits each week, month, and year, as well as overall bounce rates (i.e., percentage of visitors leaving the web site from the home page without performing a search or clicking on any page links). Google Analytics uses tracking cookies to identify unique visitors. Unique searches were recorded along with subsequent profile lookups for scientists, grants, or departments, or co-author or co-investigator network visualization for publications or grants. To detect popular topics searched by CUSP users, we applied content mining to all search terms using Automap (http://www.casos.cs.cmu.edu/projects/automap).
Using Google Analytics timestamp data, we applied time series analysis to address research questions two and three, i.e., to examine the differences of usage trends between different access modes (restricted vs. open versions) and in different years (2012 and 2013). We created time series models using Weka 3.7.9, an open-source machine learning system, to compare the temporal trends between the two access modes and the two time periods. On this basis, we applied the multi-trend regression algorithm based on support vector machine using Weka’s SMOreg function to build a trend model for each of the following time periods: restricted access, open access, open-access in year 2012, and open-access in year 2013. We compared the trend model of restricted access with that of open access. We evaluated root-mean-square error (RMSE) to quantify differences between the two access modes and the two time period models, where higher RMSE value indicated more differences. The unit of analysis for research question two was a 100-day period of each access mode: December 2, 2011 through March 11, 2012 for restricted access and March 20, 2012 through June 28, 2012 for open access. For research question three, comparing usage in different time periods during the open access phase, the unit of analysis was a 6-month period to avoid the influence of seasonal changes: March 20, 2012 through September 19, 2012 for open access 2012 and March 20, 2013 through September 19, 2013 for open access 2013.

Results

General Usage of CUSP

During the 21 months from December 2, 2011 through September 19, 2013, 4,974 unique users from 88 countries used CUSP, with a total of 8,492 visits, 28,196 page views, an average of 3.3 pages or 3-minute stay per visit, and a bounce rate of 56%. Half of the visitors (51%) landed directly (4,329 visits) by clicking the CUSP link in the signature section of emails that they received (e.g., faculty signatures that include a link to CUSP). The others (44%) landed through referrals such as the web sites for Columbia University Medical Center, Columbia University College of Physician and Surgeons (3,757 visits), or Google search (378 visits).

Excluding the bounced visitors who left the web site immediately, approximately 60% of the remaining users (2,058 visits) spent between one and three minutes on CUSP, with 20% of them (678 visits) staying for more than 10 minutes. In terms of access methods, 36% of visits came through Intranet within the Columbia University Medical Center, and approximately 7% of users accessed CUSP using a mobile device, such as a smart phone, iPad, or tablet. In terms of the temporal trends of usage, the number of unique visitors spiked to 214 per day between March 20, 2012 and March 22, 2012, when the open-access version of CUSP was released. Afterwards, the number of unique visitors per day stabilized at 10 to 30 during weekdays.

CUSP allows searches for scientists using person names or topics that appear in their publications or grants. A total of 10,210 searches were performed by 49% of all the unique visitors. Most of top 10 searches used person names directly (60%), while the rest (40%) used health topics to retrieve the collaboration network related to the topic. Content mining identified the following 20 most popular topics in Table 1.

Table 1. Frequencies of The Top Search topics

| Topic                     | Frequency |
|---------------------------|-----------|
| Diabetes                  | 212       |
| Obesity                   | 137       |
| Cancer                    | 112       |
| Irving Institute          | 75        |
| HIV                       | 71        |
| Prevention                | 64        |
| Heart                     | 62        |
| Informatics               | 59        |
| Biomedical                | 47        |
| Cell                      | 46        |
| Health disparities        | 43        |
| Comparative-effectiveness | 42        |
| Ear                       | 40        |
| community                 | 35        |
| Data                      | 34        |
| Cardiology                | 34        |
| Genetics                  | 29        |
| Breast cancer             | 29        |
| Pediatric                 | 28        |
| Nursing                   | 28        |

Figure 1 is an action transition graph generated by ORA (http://www.casos.cs.cmu.edu/projects/ora) using the page access statistics. It illustrates the patterns of the action transitions among the six frequently visited web pages on CUSP. The colors, blue or red, indicate two types of activities grouped by structural similarity in the network as detected by a subgrouping algorithm (CONCOR), which seeks to identify structurally equivalent nodes. The edge width indicates the frequency of transitions between each pair of pages. Users starting from person profile (blue on the top) pages usually reach the publication and grant pages related to a person, whereas users searching by topic (red in the center) reach topic-related grant, publication, or network pages that list names of associated investigators.
**Figure 1.** Action transition graph in CUSP (links and nodes sized by path frequency)

**Change of CUSP Usage Patterns from Restricted Access (Columbia identifier only) to Open Access**

Table 2 compares the usage statistics for the two access modes of CUSP during different time periods. Four times as many unique users visited in the open access mode than in the restricted mode (846 vs. 170). Each user spent 72% less time (11.7 vs. 3.3 minutes) and viewed fewer pages/visit in the open access mode than in the restricted one. A 44% higher bounce rate (38% vs. 55%) was observed in the open access mode. In terms of access, the number of both Intranet and Internet users increased (435% and 52%, resp.). Mobile device usage increased 523% (from 201 to 1,076) after open access.

While both restricted and open-access show a stable number of unique users over time, open access shows a one time spike at the starting point (Table 2). Time series analysis shows that the trends of bounces, search refinement and search depth of the two access modes are similar (RMSE<2). Comparing the two modes, trends of visits, unique visitors, and unique searches are only moderately different, with RMSE ranging between 28 and 55. In contrast, patterns of user time spent differ markedly between two modes, with RMSE > 1,000. Using a multi-regression modeling analysis of time series, we were able to plot a forecast model for the number of daily CUSP users of each access mode (Figure 2). According to the time series models of each access mode in Figure 2, the number of daily users of the restricted access version of CUSP continuously decreased. In contrast, the number of daily users of the open access version has steadily increased.

**Figure 2.** Forecast of the number of daily CUSP users of different access modes using time series analysis
Table 2. CUSP usage in different access modes over time

| Access Mode                  | (unit: 100 days) | Year (unit: 6 months) | Delta (%) | RMSE | Delta (%) | RMSE |
|------------------------------|------------------|-----------------------|-----------|------|-----------|------|
|                              | Restrict access  | Open access           |           |      | 2012 Open | 2013 Open |
| Visits                       | 634              | 1,735                 | 174       | 30   | 2,607     | 2,244 | -14 | 395 |
| New                          | 170              | 820                   | 382       | 2    | 1,220     | 1,631 | 34  |      |
| Returning                    | 464              | 915                   | 97        | 1    | 1,387     | 613   | -56 |      |
| Unique visitors              | 170              | 846                   | 398       | 27   | 1,245     | 1,706 | 37  | 178 |

Engagement

| Duration(min:sec) | Restrict access | Open access | Delta (%) | RMSE | 2012 Open | 2013 Open | Delta (%) | RMSE |
|-------------------|-----------------|-------------|-----------|------|-----------|-----------|-----------|------|
| 11:40             | 3:20            | -72         | 2,434     | 3:14 | 2:29      | -23       | 6,551     |
| New               | 4:29            | -38         | 2:49      | 2:49 | 1:19      | -53       | 1,19      |
| Returning         | 14:18           | -73         | 3:36      | 3:36 | 5:37      | 56        | 56        |
| Bounce            | 38%             | 55%         | 44        | 1    | 56%       | 57%       | 2         | 6    |
| New               | 58%             | 46%         | -21       | 47%  | 63%       | 34        |           |
| Returning         | 31%             | 63%         | 104       | 63%  | 42%       | -33       |           |
| Page/visit        | 7               | 4           | -51       | 3.4  | 2.9       | -15       | 37        |
| New               | 4.2             | 3.8         | -10       | 3.8  | 2.4       | -37       |           |
| Returning         | 8.2             | 3.2         | -61       | 3.1  | 4.4       | 42        |           |

Content

| Page views        | Restrict access | Open access | Delta (%) | RMSE | 2012 Open | 2013 Open | Delta (%) | RMSE |
|-------------------|-----------------|-------------|-----------|------|-----------|-----------|-----------|------|
| 4,514             | 6,043           | 34          | 203       | 8,973 | 6,546     | -27       | 3,482     |
| Search            |                 |             |           |      |           |           |           |      |
| Unique search     | 1,002           | 2,689       | 168       | 55   | 3,656     | 2,326     | -36       | 754  |
| % refinement      | 32%             | 46%         | 42        | 48%  | 50%       | 4         | 7         |      |
| Search depth      | 1.5             | 0.7         | -51       | 2    | 0.17      | 0.16      | -6        | 6    |

Changes of CUSP Usage Patterns in 2012 and 2013 in Open Access

Table 2 shows that although the number of returning visits decreased substantially in 2013, the numbers of new and unique users steadily increased. The users spent less time viewing slightly fewer pages per visit in 2013 than in 2012. Bounce rate remained similar in 2013 despite the substantial improvement in bounce rate among returning users. In terms of access, while Intranet users decreased, Internet users increased coming from more countries. The mobile or tablet use increased 90% in 2013. The trend graph shows that the number of unique users was higher in 2013 than in 2012. Time series analysis shows that bounce, search refinement, and search depth trends demonstrate similar patterns in 2012 and 2013 (RMSE ≤7). However, the trends related to user time spent, page views and unique search revealed completely different patterns between 2012 and 2013 (RMSE>700).

Figure 3 illustrates the overall usage flow of CUSP in 2012 and 2013. The numbers on the flow chart arrow in Figure 3 compare the number of unique searches in 2012 and 2013. While the number of CUSP users coming through institutional website and general search increased, the users from direct approaches (e.g., link in email) decreased in 2013 compared to 2012. Also Figure 3 shows that there were fewer unique searches for every page in
2013 than there were in 2012. Furthermore, this figure shows that the co-author or co-investigator networks of grants or publications were less utilized than individual profile pages. The numbers on the arrow pointing publication and grant network pages on the right corner show that less than 2% of unique searches (40 out of 2,389) reached the network visualization page, suggesting that CUSP’s visualization feature is underused.

Figure 3. Uses of CUSP in 2013 (numbers outside parentheses) and 2012 (numbers within parentheses)

Discussion

The results of this study have important implications for RNS designs to support biomedical researchers and RNS usage analysis methodology. We observed a slightly increased bounce rate and decreased time spent during the open access mode. However, the small values of RMSE calculated from trends analysis suggest that the temporal change patterns were not much different between the two access modes (Table 2). In addition, the average session duration decreased substantially from 11:40 in restricted access to 3:20 in open access. This decrease can be explained by several factors, such as an improved user interface for biomedical researchers in the open access CUSP (Figure 4) and different needs of Columbia and non-Columbia users. Future causal studies for these results are warranted.

Results of the trend models developed for this study offer general guidance for RNS design (e.g., open access). The trend models built based on the 100 days of use of each access mode in Figure 2 showed a steady increase of users in open access mode compared to the steady decrease during CUSP’s restricted access period. This implies that an open-access model benefits the sustainability of an RNS. The trend models also inform us about how long the effect of marketing (e.g. newsletters) has lasted in Figure 2. Other institutions should consider an open access RNS in order to support interdisciplinary collaboration.

Furthermore, this study presented a novel integration of analytical methods that may be useful to others. The Google Analytics approach provided us with rich time trends data. We were able to obtain log data representing more than 30 different kinds of user behaviors and characteristics. The data were easily imported into various analytical software applications for further analysis, including time series and content mining. While the methods were simple, the results are sophisticated. The time series analyzed for this study have rarely been applied to similar log file studies that use Google Analytics. This trend analysis based on machine learning algorithm is more powerful and sophisticated than the traditional statistical time series analysis using ARMA (autoregressive moving average) or autoregressive integrated moving average (ARIMA) model.

In addition, content mining using Automap provides more accurate use information related search terms than the output from Google Analytics. Google Analytics calculates frequency of key terms based on the morphology of the words. In contrast, our content mining approach allows us to aggregate semantically related terms utilizing natural language processing. Google Analytics does not recognize word variation, such as upper vs. lower case, word vs. symbol, singular vs. plural, and alternate spellings. For example, while Google Analytics considers “comparative effectiveness”, “comparative-effectiveness”, “Comparative effectiveness” and “Comparative Effectiveness” as four distinct terms, our approach recognizes them as a single concept.
This study inherits the limitations of log file analysis. While this web usage mining provides answers to who, how, where and when questions related to system use, it does not provide answers to why - why users interacted less with certain pages or why users are satisfied or not satisfied with the system. Nevertheless, the usage mining was valuable to assess the global picture of the usage patterns for hypothesis generation to guide further user modeling.

**Figure 4.** Improved search interface of CUSP and process for biomedical researchers to search for collaborators

**Conclusions**

This paper reports temporal usage patterns of our locally developed research network system, CUSP, by applying various temporal data mining methods to the Google Analytics usage log. Software used in this study is freely
available on the Internet. The software packages are easy to use by researchers without significant programming skills who need to mine usage patterns on a health-related website. Our temporal usage analytical framework allowed us to efficiently characterize and understand the temporal usage of open-access RNSs like CUSP and our results offer generalized guidance for improving the design of future RNSs.

Acknowledgments

The authors of this study were supported by grants R01 HS019853 (PI: Bakken), R01 LM009886 (PI: Weng), and UL1 TR000040 (PI: Ginsberg). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Liu B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). New York: Springer Heidelberg Dordrecht London; 2011.
2. Boland MR, Trembowelski S, Bakken S, Weng C. An initial log analysis of usage patterns on a research networking system. Clinical and translational science. Aug 2012;5(4):340-347.
3. Bracke PJ. Web usage mining at an academic health sciences library: an exploratory study. J Med Libr Assoc. Oct 2004;92(4):421-428.
4. Clifton B. Advanced Web Metrics with Google Analytics. Indianapolis, Indiana: John Wiley & Sons, Inc; 2012.
5. Bisgaard S, Kulahci M. Time Series Analysis and Forecasting by Example. Hoboken, New Jersey: Wiley; 1 edition 2011.
6. Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK. Improvements to the SMO algorithm for SVM regression. Neural Networks, IEEE Transactions on. 2000;11(5):1188 - 1193.
7. Saigal S, Mehrotra D. Performance comparison of time series data using predictive data mining techniques. Advances in Information Mining. 2012;4(1):57-66.