Did You Blog Yesterday? Retention in Community Blogs

Imrul Kayes*, Xiang Zuo*, Da Wang**, Jacob Chakareski+
*Computer Science and Engineering,University of South Florida,Tampa, Florida, USA
**Electrical and Electronic Engineering, HuBei University of Technology,Wuhan, China
+Electrical and Computer Engineering, University of Alabama, Tuscaloosa, Alabama, USA
{imrul,xiangzuo}@mail.usf.edu, wangda222@126.com, jacob@ua.edu

ABSTRACT

We ask what factors cause a blogger to continue participating in the community by contributing content (e.g., posts, comments). We crawled a sample of blogger profiles (contributed 91% posts) from a popular community blogging platform "Blogster". Our results show that the male and aged (senior) bloggers, who face fewer constraints and have more opportunities in the community are more retained than others. Other bloggers pay a high degree of attention to these retained bloggers through implicit (reading posts) and explicit (writing comments) interactions. We have also found that a blogger has higher retention if her friends have also higher retention and a strong social tie reduces retention imbalance between two blogger friends. However, we found that a blogger's network age (e.g., how long ago she joined) has no effect on her retention.

Keywords

Community Blogs, Retention, Bloggers

1. INTRODUCTION

Community blogging is a medium for publishing daily journals, expressing opinions or ideas, and sharing knowledge. Blogging has a high impact on marketing, shaping public opinions, and informing the world about major events from a grassroots point of view. However, turnover in online blogging is very high, with most people who initially join and start contributing to the community, failing to contribute in the long run. Existing research, based on content analysis [2] and interviews [3], provides an excellent insight into a user's motivation for joining a blogging community. However, the question of why bloggers continue participating, in other words why some bloggers have higher retention, is not well researched.

In this work, we examine the factors that cause a blogger to continue participating on the platform. In doing so, we first crawled a sample of bloggers' profiles from an online community blogging platform Blogster. These users contributed about 91% of posts in the community, and hence are a good candidate for a representative sample of the bloggers in Blogster. We propose five categories of variables: network-specific (e.g., centralities, clustering coefficient), user activity specific (e.g., posts, comments, photos, network age), physiological (e.g., age, gender), interactional (e.g., blog traffic, other users' comments) and relational (e.g., social tie strength, friends retention) to understand which factors contribute to continued participation or higher retention and to what degree. Finally, we put together the variables to predict a blogger's retention in Blogster. We found that our model fits well the data and predicts retention with $Adjusted\ R^2 = 0.84$.

2. DATA COLLECTION

We used the Metropolis-Hastings Random Walk (MHRW) algorithm [1] to crawl the Blogster network. The MHRW algorithm is capable of obtaining a uniform sample (or more generally a probability sample) of OSN users [1]. We scraped HTML pages of $17,436$ users, who form a social graph with $72,907$ edges and 17 connected components. The largest connected components has $14,323$ nodes and $64,888$ edges. Blogster has a public post counter, where it shows the number of posts bloggers have already contributed. Our crawled bloggers contributed $329,114$ posts out of $362,123$ posts on Blogster, as seen by the post counter. So, bloggers from our dataset contributed about 91% of total posts.

3. RESEARCH QUESTIONS

We have several hypotheses that are related to the following research questions:

1. What variables predict high retention?
2. How well do these variables predict user retention?

In Twitter, Java et. al. [2] found that active users have higher retention. So, simply putting this forward in the context of Blogster, retention is how active a blogger is in the network, in other words the user's engagement or participation with the community. One can measure a user's participation in the network by continuously monitoring her activities. In Blogster, participation is measured by points. To encourage participation Blogster has a point system. The number of points a user gets depends on the specific actions she takes [1]. We take the points explicitly stated on a blog-

BigDataScience '14 August 04 - 07 2014, Beijing, China
Copyright 2014 ACM 978-1-4503-2891-3/14/08.
http://dx.doi.org/10.1145/2640087.2644178 ...$15.00.

[1] www.blogster.com/help#earnpoints

ger's profile as an indication of her retention. Users with higher points have higher retention.

We categorized the predictor variables of retention into five categories: network metrics specific variables (clustering coefficient, degree, betweenness, closeness, pagerank, and communicability centrality); user activity specific variables (# posts, # comments, # photos, network age); user physiology oriented variables (age, gender); interactional (blog traffic, other users' comments); relational (social tie strength, friends retention).

4. PREDICTING RETENTION

A multiple regression model to predict retention (as reflected by points) using only all activity specific variables shows that activities can highly predict (96%) retention. In fact, in Blogster activities are most contributing factors of points. For example, writing a blog post contributes 15 points, writing a comment per single blog post contributes 2, adding a photo as avatar gives 15 points, deleting a blog post reduces points by 15 and deleting a comment reduces points by 2. So, we plan to exclude all activity specific variables in our prediction. Our goal is to investigate to what extent variables that are not related to point calculation can predict user retention.

As an input processing step, first we rank users based on each centrality metric. Specifically, each centrality metric assigns each node a score that can be used to order users in decreasing order of importance (according to that centrality). This allows each blogger to receive a rank according to each centrality metric: the first ranked blogger will be the most central one, the last ranked will be the one with the lowest centrality score. Bloggers having the same centrality score are given the same rank.

	Estimate (β)	Std. Error	t value	Pr($>|t|$)
(Intercept)	3.66e+03	1.86e+02	19.62	$< 2e-16$ ***
DegreeRank	-1.92e+01	1.07e+00	-17.98	$< 2e-16$ ***
Comm.Rank	-1.57e-02	4.25e-03	-3.68	0.000233 ***
CC	-8.71e+01	4.06e+01	-2.14	0.032081 *
BlogTraffic	3.62e-02	8.21e-04	44.06	$< 2e-16$ ***
UserComments	1.02e+00	1.59e-02	64.61	$< 2e-16$ ***
Age	5.03e+00	7.29e-01	6.90	$5.45e-12$ ***
Gender	-9.38e+00	1.86e+01	-4.50	0.001257 **
AvgFriendsRet	1.53e-04	2.04e-03	3.08	0.021513*

Table 1: The results of a multiple linear regression with points as the dependent variable. Significance codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.**

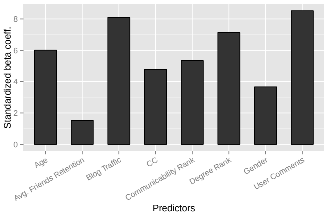

Figure 1: Standardized beta coefficients (β) show the relative effect sizes that each input variable has on points.

We considered all explanatory variables other than activity specific variables to build a multiple regression model to predict retention (as reflected by points). However, not all predictors are significant in explaining the variability of points. So, our regression went through a variable selection process. We used an "all possible regression" approach to select model predictors. The all possible regression approach considered all possible subsets of the pool of explanatory variables and found the model that best fits the data according to adjusted R^2. Finally, we considered the model that yielded maximum adjusted R^2. The model is the following:

$$\begin{aligned} Points_i = \alpha + \beta_1.DegreeRank_i + \beta_2.CC_i + \\ \beta_3.CommunicabilityRank_i + \beta_4.BlogTraffic_i \\ + \beta_5.UserComments_i + \beta_6.Age_i + \\ \beta_7 Gender_i + \beta_8 AvgFriendsRet_i + \epsilon_i \end{aligned} \tag{1}$$

The results of the regression are shown in Table 1. The adjusted R^2 of the model is 0.837, which implies the model can explain 83.7% of variation around points. The unstandardized β coefficients in Table 1 are useful in that they can be directly interpreted according to the native units of each predictor: for each one unit change in the predictor variable, the count of the response variable (points) is expected to change by the respective *beta* coefficients (all else being equal). As expected higher degree rank, communicability rank and clustering coefficient means lower points with corresponding β are $-1.92e+01$, $-1.57e-02$ and $-8.71e+01$ respectively. However, the more web traffic and other users' comments a blogger gets, the more points she can expect with $\beta = 3.62e-02$ and $\beta = 1.02e+00$ respectively. Also, being female suggests less points $\beta = -9.38$ and aged user might expect higher points with $\beta = 5.03$. Furthermore, if a blogger's friends has higher retention, we might expect her retention higher with $\beta = 1.53e-04$.

While β coefficients are valuable for a broad range of prediction and forecasting purposes, we are also interested in comparing the relative impact of each predictor. We report the standardized beta β coefficients in Figure 1. From the figure we observe the number of other users' comments, web traffic and the degree rank are the most influential or significant predictors. The rest of the predictors can be serialized from the most significant to the least significant ones as age, communicability rank, clustering coefficient, gender and average friends retention.

5. SUMMARY

In this study of blogger retention, we found that male and aged (senior) bloggers, who occupy central positions, have low clustering coefficient in the blogging community and have friends with higher retention are more retained in the community than others. These bloggers also get higher attention from others as reflected by higher explicit and implicit interactions from others. Our work can be used as a foundation for further study of community bloggers retention. System developers of community blogs could also leverage results of this paper and build retention aware community blogs.

6. REFERENCES

[1] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Practical recommendations on crawling online social networks. *Selected Areas in Communications, IEEE Journal on*, 29(9):1872–1892, 2011.

[2] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[3] B. A. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz. Why we blog. *Communications of the ACM*, 47(12):41–46, 2004.