Why leave Wikipedia?

Poster Submission

October 29, 2007

1 Introduction

Some user-contributed content (UCC) applications, such as Yahoo!Answers, Wikipedia, and YouTube have drawn much media attention. Various reasons motivate the tremendous contributions to a few UCC systems so far. Existing literature has uncovered many factors that affect users' decisions to *become* contributors [Bryant et al., 2005], to *continue* contributing [Nov, 2007], and to *increase* contribution [Chen et al., 2007], but none of them pay attention to why active contributors decide to *stop* contributing. The exit of active contributors may affect quantity and quality of content provision on UCC systems.

Indeed, preliminary evidence shows that there is a reason to worry about the long term sustainability of some systems. In 2000, Adar and Huberman's study on the Gnutella network showed that there was high level of free-riding on this network. Five years later, Hughes et al. [2005] find on the same network free-riding has gotten worse: the percentage of users who do not share any files has increased from 66 % to 85%. In the mean time, Andrew Lih, one active Wikipedia research/contributor also blogged about its growth rate slowing down dramatically in late 2006 [Lih, 2007]. To what extent is this due to contributors leaving the system? If so, who left? And for what reasons? Answers to these questions are useful to UCC system designers for determining the impact of contributors leaving or for devising mechanisms to prevent them from leaving.

In this study we focus on one system, Wikipedia, and analyze why some editors stop contributing. We chose Wikipedia for a number of reasons. First, it has a large number of active contributors. By mid-2006 over 10,000 editors had made more than 100 edits [Kittur et al., 2007]. Second, Wikipedia maintains a detailed record of its contributors' activities, and shares its database online. [1]

2 Method

We will conduct statistical analysis on a dataset we obtained from the English Wikipedia on July 16, 2007. This dataset contains the full edit history of every article and user page on Wikipedia. We take a simple random sample of all the registered editors. In

[1] See http://en.wikipedia.org/wiki/Wikipedia:Database_download.

Table 1: Predictors of Wikipedians' Decisions to Exit.

	Name
Independent variables	creator dummy
	preserver dummy
	destroyer dummy
	D_Ontime
	D_Deled
	work intensity
	article stability
Controls	total lifetime number of edits

our study, an editor is said to ”leave Wikipedia” if they do not edit for three months or longer. The choice of three months as a cut-off is arbitrary. In the following sections we explain how we operationalize other variables, and state our hypotheses.

2.1 Lifespan of a Wikipedian

First, we ask how the lifespan of a Wikipedian is distributed. We estimate this distribution using a survival analysis with truncated data — treating the active Wikipedians' data points as truncated[2] — while assuming a parametric distribution function such as lognormal or Weibull. This distribution provides a baseline for exit behavior. We can use this distribution to estimate, for example, the effects of changes in exit timing.

2.2 What predicts a Wikipedian's departure?

Next, we identify factors that predict a Wikipedian's departure and estimate their effects. We analyzed four hypothesized predictors: roles played in the system, peer feedback, work intensity, and article stability. The effects of these predictors will be estimated simultaneously in a logit regression, controlling for the editor's total number of edits. Table 1 lists our independent and control variables.

For each observed editor, we define three "decision periods": *lifetime*, *last week*, and *last month*. The lifetime of an editor begins with her first edit and ends with her last edit in our dataset. If an editor is deemed as having "left" Wikipedia, her *last week (month)* is the last week (month) of her participation, not the last week (month) in our dataset. For a currently active editor, the *last week (month)* refers to the last week (month) in our dataset.

2.2.1 Roles

We identify three different roles among Wikipedians: creator, preserver, and destroyer. An editor plays one or more of these roles depending on the type of edits she contributed.

[2] That is, their observed lifespans are shorter than their actual lifespans due to the cutoff date of observation.

We classify all edits into four mutually exclusive types: creation, reversion, deletion, and damage. Reversions are easy to identify: if the texts are exactly the same as an earlier version, judged by the MD5 checksums of the texts, it is a reversion. With the help of techniques introduced by Priedhorsky et al. [2007], we can identify damages. First, edits that are reverted in future are identified as will-be-reverted (WBR) edits. Damages are WBR edits "where the future reverts edit comment suggests either (a) explicit intent to repair vandalism or (b) use of revert-helper tools or autonomous anti-vandalism bot activity." If an edit is not classified as a reversion or a damage, it is classified either as a creation or as a deletion. If an edit increases the total number of words in the text or it adds new words into the text, it is a creation. On the other hand, if an edit decreases the total number of words in the text and no new words are added, it is a deletion. Suppose editor i has edited w_i words in her lifetime, within which w_{ci} were created, w_{ri} were involved in her reversions, w_{di} were deleted, and w_{mi} were damages. We calculate the proportion of her creation as follows,

$$p_{ci} = \frac{w_{ci}}{w_{ci} + w_{ri} + w_{di} + w_{mi}}; \tag{1}$$

and her proportion of reversions and deletions as follows,

$$p_{rdi} = \frac{w_{ri} + w_{di}}{w_{ci} + w_{ri} + w_{di} + w_{mi}}; \tag{2}$$

and lastly, her proportion of damages as follows,

$$p_{mi} = \frac{w_{mi}}{w_{ci} + w_{ri} + w_{di} + w_{mi}}. \tag{3}$$

Let p_c denote the population mean of p_{ci}, p_{rd} the population mean of p_{rdi}, and p_m the population mean of p_{mi}. If $p_{ci} > p_c$, editor i is labeled as a creator. If $p_{rdi} > p_{rd}$, editor i is labeled as a preserver. And if $p_{mi} > p_{mc}$, editor i is labeled as a destroyer. We do not have strong prior hypotheses regarding how editor roles affect departure. Our findings will inform us about the effect of exit behavior on some aspects of the quality of UCC contributions.

2.2.2 Recent peer feedback

It is rewarding to see ones' own edits persist. It may also be disappointing to see ones' own edits being removed by others. How other editors react to one's edits serve as feedback, and people seek self-satisfaction from peer feedback [Ryan and Deci, 2000]. In addition, Wikipedia makes it easy to watch peers' feedback by providing a watchlist function, so that a registered editor can monitor changes made on any page, certainly including the pages that she edits.

We define two variables, *Ontime*— the number of minutes that a Wikipedian's edits persist, and *Deled* — the number of times a Wikipedian's edits get deleted. We hypothesize that if $Ontime$ increases for an editor, she is rewarded and is motivated to contribute more, hence is less likely to quit. On the other hand, if $Deled$ increases, she is more likely to leave. For editor i, we denote the changes in these variables by D_Ontime_i and D_Deled_i, and derive them as follows,

$$D\_Ontime_i = Ontime_i \text{ (last week) } - \; Ontime_i \text{ (week before last week)} \quad (4)$$

$$D\_Deled_i = Deled_i \text{ (last week) } - \; Deled_i \text{ (week before last week)} \quad (5)$$

Our hypotheses are that the probability of leaving decreases in D_Ontime_i, and increases in D_Deled_i.

2.2.3 Last month work intensity

Within the Wikipedia community some speculate that getting burnt-out causes some editors to quit [Lih, 2007]. Thus those editors who have left may have worked harder than usual right before leaving. We propose to use the following formula to describe the work intensity in the last month of editor i,

$$\text{Last month work intensity}_i = \frac{\text{number of edits in the last month}}{\text{lifetime average number of edits per month}} \quad (6)$$

We hypothesize that the higher a Wikipedian's work intensity, the more likely that she will leave.

2.2.4 Last week article stability

Some Wikipedians are only interested in editing a particular set of pages. When these pages reach a stable stage, they may find little motivation to continue. We identify the articles an editor cares about by looking at how much she has contributed to them. Suppose editor i has contributed to n articles in total, with e number of edits to each of these articles on average. We define all articles which have received more than e edits from editor i as the set that i cares about. The stability of an article is measured by the number of edits it receives in the last week of editor i. The stability of the set of articles is the average stability of each article in the set. We hypothesize that the more stable the articles that an editor cares about, the more likely that this editor will stop contributing.

Acknowledgments

This section is blanked out for blind peer review purposes.

References

Eytan Adar and Bernardo A. Huberman. Free riding on Gnutella. *First Monday*, 5, 2000.

Susan L. Bryant, Andrea Forte, and Amy Bruckman. Becoming wikipedian: transformation of participation in a collaborative online encyclopedia. In *GROUP '05:*

Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work, pages 1–10, New York, NY, USA, 2005. ACM. ISBN 1-59593-223-2. doi: http://doi.acm.org/10.1145/1099203.1099205.

Yan Chen, Max Harper, Joseph Konstan, and Sherry Xin Li. Social comparisons and contributions to online communities: A field experiment on movielens, 2007. Working paper.

Dianel Hughes, Geoff Coulson, and James Walkerdine. Free riding on gnutella revisited: the bell tolls? *Distributed Systems Online*, 6(6), June 2005. doi: 10.1109/MDSO.2005.31.

Aniket Kittur, Ed Chi, Bryan A. Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007)*, April 28 - May 3 2007.

Andrew Lih, 2007. URL `http://www.andrewlih.com/blog/2007/06/28/wikipedia-plateau/`.

Oded Nov. What motivates wikipedians, 2007. URL `http://opensource.mit.edu/papers/Nov_Wikipedia_motivations_opensource.mit.edu.pdf`. Accepted for publication in the Communications of the ACM.

Reid Priedhorsky, JIlin Chen, Shyong (Tony) K. Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, destroying, and restoring value in wikipedia. In *GROUP '07: Proceedings of the 2007 international ACM SIGGROUP conference on Supporting group work*, New York, NY, USA, 2007. ACM.

Richard Ryan and Edward L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55 (1):68–78, 2000.