Investigating Social Interaction Factors for Newcomer Retention in Wikipedia Teahouse

Moyan Zhou
University of Minnesota
Minneapolis, Minnesota, USA

Ruixuan Sun
University of Minnesota
Minneapolis, Minnesota, USA

Loren Terveen
University of Minnesota
Minneapolis, Minnesota, USA

Keywords: Wikipedia Teahouse, Socialization, Newcomer Retention, Survival Analysis, Social Interactions

Introduction

Newcomers benefit online communities: they bring innovative and diverse ideas and fill in the gaps created by members who leave the community (Morgan and Halfaker, 2018). However, onboarding newcomers and retaining them is challenging (Halfaker et al., 2013), as newcomers tend to drop out even after their first session (Karumur et al., 2016). Thus, online communities dedicate efforts to understand and to promote newcomer retention, one approach being socialization. For example, Wikipedia Teahouse, a Q&A forum, encourages social interactions between newcomers and more experienced editors. Prior work has found that the Teahouse supports newcomer retention (Morgan et al., 2013).

The positive effect of Wikipedia Teahouse hints relationships between socialization and newcomer retention, which is essential to the success of online communities. Thus, to further investigate this relationship, we ask: how do social interactions with other community members affect newcomer commitment in peer production platforms? In the context of Wikipedia Teahouse, we ask: how does Teahouse interactions affect newcomer retention? Answering this question will provide insights for community members on how to interact with newcomers, potentially extending to other forms of social interactions such as talk page discussions.

In this workshop paper, we consider Wikipedia Teahouse as a case study, and unpack how social interactions with more experienced community members affect newcomers' commitment to Wikipedia. We aim to understand and thus enhance socio-technical interventions designed to retain newcomers by socialization, and to assist newcomer decline in the long term. By identifying specific factors that correlate with newcomers' retention in English Wikipedia, we contribute to Wikipedia through the following research questions:

RQ: *What* and *how* do social interaction factors affect newcomer retention in Wikipedia?

Methods

Hypothesized Variables In the Teahouse, newcomers who ask questions are referred to as "guests," while more experienced editors who answer the questions are referred to as "hosts." We follow this convention. And we hypothesize variables based on prior work in knowledge sharing communities, member motivation and participation. We divide these variables into two broad groups, shown in Table 1.

Data Collection and Filtering All Teahouse Q&A records are stored in publicly available archives. After extracting and processing wikitext, we have labels for each guest, host and follow up. We also pattern match title, content, timestamp, and extract contribution and registration data for every guest. Our validation shows 98% agreement in these information among 50 randomly selected samples. Moreover, we filter newcomers as guests who made no more that 100 edits in all namespaces within 30 days of registration (Morgan et al., 2013). We end up with 16007 Q&A interactions in the final dataset.

Survival Analysis We conduct survival analysis in a Cox Proportional-Hazards model. Survival analysis shows the relationships between predictor variables and event of interest, taking time to event into consideration. When the variables are normalized, the hazard ratio represents the change in the risks of withdrawal as a predictor variable increases by one unit in standard deviation (Wang et al., 2012).

Withdrawal and **Time**: We define withdrawal as 90-days of inactivity (i.e., no contributions to any namespace) on Wikipedia (Yu et al., 2017), which is censored to 0 indicating the user could still be alive if their last edit is within 90 days from the date of collection. Then we calculate the time duration in months from their question on the Teahouse to their last edit on Wikipedia.

Results

Model 1 shows the significant relationships between control variables and withdrawal. More specifically, one unit increase in Std Dev of *contributions* and that of *registration* is associated with a 26.6% decrease and a 7.3% decrease in the risk of withdrawal.

Model 2 reveals the significant effects of sentiment of answers (*tone*, *analytic* and *authentic* but not *clout*) on withdrawal, in addition to control variables. The result implies that newcomers are respectively 2.7%, 1.9%, and

3% more likely to stay in Wikipedia if they receive answers that score one-unit Std Dev higher than average in *tone*, *analytic* and *authentic*.

Model 3 shows that *numpolicies* influences newcomer withdrawal. One Std Dev increase in the number of policy links mentioned in the first answers is related with a 10.9% increase in the likelihood of withdrawal from Wikipedia.

Model 4 reports that *nummessages*, *answerlength*, and *waitingtime* affect newcomer withdrawal. The risk of withdrawal for newcomers decreases by 5.3%, 5.5%, 2.6% as the number of interactions, word count of the first answer, and waiting time increases by one Std Dev.

Discussion/Conclusions

Interpretation of Results Our results confirm that *Wikipedians are both born and made*. On one hand, Wikipedians are born. The significant and strong effects of control variables indicate that people' intrinsic characteristics play a major role in their Wikipedia careers, aligning with prior work on the notion about "Wikipedians are born" (Panciera et al., 2009). On the other hand, Wikipedians also can be made. The social interactions on the Teahouse matter and significantly reduce the risk of withdrawal. More interactions with experienced editors foster social ties and engagement. Longer answers provides more details that could make answers easier and clearer to understand for newcomers. Longer response time for the first answer is associated with higher retention, possibly due to better quality when hosts take longer to phrase their answers, but further investigation is needed. The results also suggest general characteristics of helpful responses to newcomers: these responses are in a positive tone, provide necessary details and evidence, tend to be in casual and conversational style, and avoid referring to Wikipedia policies with direct links.

Design Implications There are two design implications from our results. First, our findings serve as the basis for guidelines to Teahouse hosts that remind them of what makes a helpful response; the guidelines could be included as checklist in hosts' user interface as hosts create and edit their responses. Second, rather than relying solely on hosts to apply guidelines, the guidelines could be embodied in one or more support tools. One such tool could use generative AI techniques to take a host's response, along with a suitable prompt to make it (for example) "more positive, more conversational, and more informative." Helpful responses can be produced effectively through the collaboration between experienced editors and intelligent support tools.

Beyond committed newcomers New Wikipedia editors who post questions on the Teahouse already differ from typical new editors: they make more edits, decide to ask questions, and engage in help-seeking behaviors. For a typical newcomer, the relative effects of identified factors may be different. Thus, future study should consider more a general population of newcomers.

More complicated models Additional analysis can be useful. For example, analyzing followups from the newcomers: do they indicate appreciation (suggesting satisfaction) or confusion (dissatisfaction)? Another potential future work lies in exploring behaviors of editors who drop out and then return to Wikipedia.

References

[Halfaker et al.2013] Aaron Halfaker, R Stuart Geiger, Jonathan T Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, 57(5):664–688.

[Karumur et al.2016] Raghav Pavan Karumur, Tien T Nguyen, and Joseph A Konstan. 2016. Early activity diversity: Assessing newcomer retention from first-session activity. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 595–608.

[Morgan and Halfaker2018] Jonathan T Morgan and Aaron Halfaker. 2018. Evaluating the impact of the wikipedia teahouse on newcomer socialization and retention. In *Proceedings of the 14th International Symposium on Open Collaboration*, pages 1–7.

[Morgan et al.2013] Jonathan T Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. 2013. Tea and sympathy: crafting positive new user experiences on wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 839–848.

[Panciera et al.2009] Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians are born, not made: a study of power editors on wikipedia. In *Proceedings of the 2009 ACM International Conference on Supporting Group Work*, pages 51–60.

[Wang et al.2012] Yi-Chia Wang, Robert Kraut, and John M Levine. 2012. To stay or leave? the relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 833–842.

[Yu et al.2017] Bowen Yu, Xinyi Wang, Allen Yilun Lin, Yuqing Ren, Loren Terveen, and Haiyi Zhu. 2017. Out with the old, in with the new? unpacking member turnover in online production groups. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–19.

Rationale	**Hypothesized Variable(s)**	**Note**
Individual properties are likely to affect guests' retention.	*contributions* (control)	Number of contributions before the guest posts a question on the Teahouse.
	registration (control)	Amount of days the guest has been in Wikipedia after registration before they ask their first question on the Teahouse.
Social interactions with more experienced editors are likely to affect guests' retention, including structural and content factors such as sentiment.	*nummessages**	The number of messages in the conversation thread.
	*answerlength**	Word count of the first answer.
	*waitingtime**	Time the guest wait for the first answer in minutes.
	*numpolicies**	Number of Wikipedia policy links in the answer.
	tone, *analytic*, *authentic*, *clout* (summary variables from LIWC analysis)	*tone* summarizes the differences between words with positive emotions (e.g. happy, love, nice) and words with negative emotions (e.g. hate, hurt, ugly) in the text; *analytic* tells the complexity of writer's thoughts; *authentic* shows the degree of truthfulness or honesty in terms of self-presentation in the answers; *clout* refers to leadership, confidence and certainty reflected from the text.
*: Only first answers considered due to its strongest effect to withdrawal among all answers.		

Table 1: Summary of the results

	Descriptive		Model 1		Model 2		Model 3		Model 4	
	mean	std dev	HR	p-value	HR	p-value	HR	p-value	HR	p-value
contributions	7.27	13.80	0.734	***	0.735	***	0.739	***	0.739	***
registration	4.60	7.00	0.927	***	0.927	***	0.930	***	0.931	***
tone	60.28	31.99			0.973	**	0.969	***	0.968	***
analytic	54.34	29.89			0.981	*	0.982	*	0.978	*
authentic	28.96	29.03			0.970	***	0.978	*	0.976	*
clout	60.22	32.39			1.015	0.095	1.015	0.086	1.014	0.124
numpolicies	1.23	1.67					1.109	***	1.106	***
nummessages	3.83	2.38							0.947	***
answerlength	68.29	66.15							0.945	***
waitingtime	88.90	378.38							0.974	*
*: p-value < 0.05, **: p-value < 0.005, ***: p-value < 0.0005;										
After newcomers asked their first questions on the Teahouse, about 50% of them stayed after the first day, 20% of them stayed after a year, 10% stayed after 3 years and above.										

Table 2: Descriptive data and results from survival analysis