Predicting Long-Term Activity in Online Writing Communities: A Quantitative Analysis of Amateur Writing

Boot, P.

published in
Supporting Digital Humanities (Copenhagen 17 - 18 November 2011). Conference Proceedings
2011

document version
Publisher's PDF, also known as Version of record

Predicting Long-Term Activity in Online Writing Communities

A Quantitative Analysis of Amateur Writing

Peter Boot

Huygens Institute for the History of the Netherlands
PO Box 90754, 2509 LT Den Haag
The Netherlands
peter.boot@huygens.knaw.nl

Abstract

Online writing communities are web sites where amateur writers gather to publish and discuss their poems and stories. In this paper I will look at a Dutch-language example of an online writing community, Verhalensite. I will demonstrate the feasibility of statistical analysis of online writing communities by investigating what data about members' initial activity on the site, and the response they receive from fellow members can predict about their long-term careers. The paper briefly considers the implications of this sort of research for literary studies and the humanities research infrastructure.

Introduction

Online writing communities are web sites where amateur writers gather to publish and discuss their poems and stories. A large number of such sites exists. In the Netherlands only, they number at least 30. The implications of these sites for literary research are profound (Boot, 2011). As the sites contain both the texts and the comments they have drawn, they offer a wealth of empirical material on literary evaluation. Researchers with a sociological interest (Verdaasdonk, 1985) can count reactions and often ratings. And as the full texts of the works are also electronically available, online writing communities also facilitate quantitative stylistic studies. A process that used to be intractable for lack of quantifiable data is becoming amenable to study because of the existence of these online writing communities.

In this paper I will look at a Dutch-language example of an online writing community, Verhalensite (www.verhalensite.com). I will demonstrate the feasibility of statistical analysis of online writing communities by investigating what data about members' initial activity on the site can predict about their long-term careers.

Verhalensite and its Visitors

Verhalensite was an online writing community established in November 2001. Its number of members was 2685. The site ceased to operate in April 2011. The decision to discontinue the site was certainly not taken for lack of success: there are 60086 works in the download of the site that the paper is based on.

While the site's raison d'être is the exchange of works and comments, for many of its visitors the site also fulfils an important social role. Much of the comment is social rather than evaluative in nature, expressing encouragement or personal appreciation. The site has brought about friendships and marriages and many members have participated in yearly meet-ups.

Not much is known about who these site members are. Members have profile pages, but many of these are empty and the information that is provided is in free text. There is no way to get at trustworthy data about demographic properties such as sex, age and nationality. Having said that, manual inspection of the profile pages of the 250 most active members (responsible for 71% of site activity) showed this distribution for nationality and sex:

Sex		Nationality	
Male	85	Belgian	46
Female	120	Dutch	78
Unknown	45	Unknown	126
Sum	250	Sum	250

Table 1 Distribution of sex and nationality of top 250 Verhalensite members

So, more female than male active participants, and more Dutch than Belgian; but even more for whom nationality was impossible to determine. Extending this effort for less active participants would probably be pointless, as less activity tends to correlate with an empty profile page. An effort to collect age data in the same way proved unrewarding, as age is usually unreported, reported only vaguely ('over 50'), and if reported precisely might well be out of date. From the profile texts it is clear, however, that the participants include all age groups from high school to senior citizens. Students were especially active Verhalensite participants.

Texts on the site belong to one of five genres: poems (53%), stories (19%), serial stories (19%), miniatures (7%) and workshop exercises (2%). Stories are usually short. Their average length is 1087 words. Miniatures are brief texts, mostly in prose. There doesn't seem to be a firm distinction in length between stories and miniatures however. Works are also divided in 67 categories, most of them based on a theme, such as 'love', 'fear' or 'fantasy'; others are based on the occasion for which the texts were written, such as the yearly writing competitions. 'General' is (uninformingly) the most commonly employed category (27%), followed by 'love' (8%) and 'life' (7%).

Initial Activity and Long-Term Career

In the 'regular' literary world, decisions about an author's literary reputation are made right after his debut (Ekelund & Börjesson, 2002). One reason for this is that positive

reviews of a title increase critical attention for following titles (Van Rees & Vermunt, 1996), another one is the substantial effect of acquired readership on the popularity of new literary works (Verdaasdonk, 1987). In the world of fan fiction writing, Pugh (2005, p. 146) showed that interaction with readers and fellow writers is an important factor motivating writers to continue writing. It is likely this holds for other amateur writers as well.
Taken together these findings suggest that it should be possible to predict a writer's long term success on Verhalensite by looking at his initial postings and the response these postings received.

Available Data

Works and their Popularity

In June 2010 60086 works were downloaded from the site, written by 2435 authors. The works received 348627 comments. The comments in their turn were replied to 444867 times. 152050 comments were accompanied by an explicit rating in terms of one to five thumbs up.
Apart from other members' reactions and ratings, four other measures of popularity or quality are available on the site. Two reflect the behaviour of site visitors: the number of times a work was read, and the average reading time. Two other measures are awards granted by the anonymous site administration: each day one poem and one story were selected as 'poem (or story) of the day'; works could also receive an exclamation mark for being especially good or relevant. These stories are known as 'selections'.

Processing

Information about the authors (the date they joined Verhalensite) was collected from members' profile pages.

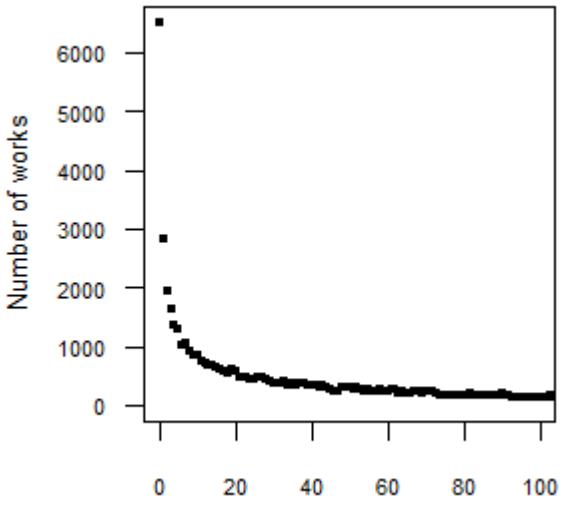

Figure 1 Number of works by week after joining the site

As appears from figure 1, the number of works published in the first weeks of membership is quite large. Based on the membership date, a subset of works was created that were published within the first four weeks of membership. It contains 12913 works, or 21% of the full collection. In order to perform an analysis of author behaviour, information about the works was aggregated at author level (n=2435). A number of variables were computed that reflect initial activity on the site and the received feedback, such as number of works published in those four weeks, number of reactions received to those works, number of reactions to other members' works, etc. In the analyses to be presented here, these are the independent variables. The dependent variables were constructed from the remaining works, those that were written after the fourth week of membership.
It should be noted the data are not as perfect as one might wish. First of all, and most importantly, stories may be removed from the site, either because users leave altogether or because they remove individual stories. This may happen for any reason: perceived lack of quality is one, but what also happens is that after a major rewrite authors remove earlier versions of a work. Sometimes authors remove works after they have been published in book form. A second limitation of the available data is that there is no one-to-one correspondence between authors and people: people may have multiple memberships, sometimes at the same time (perhaps to express multiple aspects of their personality), sometimes sequentially (perhaps having been a member in the past and returning after a few years of inactivity). Sometimes two members of the site create a third account to publish joint works. In all of these cases, it is unlikely the accounts involved will show the same usage patterns as other, 'regular' accounts. A third limitation is that there is some evidence of data having been tampered with. There are, for instance, a few members that appear to have contributed works before they were even members of the site. Over a nine year period, with software updates, procedural changes, and no doubt the occasional glitch, there will have been many reasons for manual updates to the site's database.

Variables

Of the available variables, some reflect user characteristics.

- date of membership: the day (since the creation of the site) that the user became a member of the site. Apart from actual differences over time in new user characteristics this will also reflect a self-selection effect: presumably, the authors with a short or unsuccessful publishing trajectory will tend to remove themselves from the site. We will expect older users to be more successful than younger ones.
- preference for poetry, stories, and serial stories: these variables express users' preferences for either poetry, stories or serial stories. The variables have the value 1 when more than half of a user's production falls in the given category. I expect authors of serial stories to be committed individuals who will have long publishing trajectories, though perhaps not especially popular. Poetry being usually shorter, and perhaps therefore more often written on the spur of the moment, we expect that users with a preference for poetry will as a rule have shorter careers.
- percentage of poems: number of poems as fraction of total number of works.

Other variables reflect user activity:

- number of works published on the site: may be assumed to reflect both a general urge to write and a desire to succeed on this specific site. We expect it to

have a positive correlation with number of works written over a longer span and duration of activity on the site.

- no. of reactions by author: number of reactions an author gives given to works by other site users. A high number of reactions reflects a willingness to engage with the works of others that presumably shows a desire to remain active on the site. We also expect it to be rewarded in kind (with comments by others on a user's later work).
- percentage author replies to comments: authors generally thank the commentators on their work for their reactions. This variable contains the fraction of comments that an author has replied to. We expect authors who feel at home on the site to reply more often than those who feel isolated. We also expect that replies encourage commentators to comment again on new works by the same author. Both expectations would result in a positive correlation between this variable and popularity.
- no. of author's replies to reactions. Unlike the previous variable, this variable takes into account all replies on comments, not just to comments on a user's own works. We expect effects similar to number of reactions.
- last active day: last day (of the first four weeks) that a user published a work. We expect that users who publish beyond the initial days are more likely to remain active beyond the first four weeks.

Variables that reflect work characteristics are:

- number of genres: genres practiced in the first four weeks. Publishing in multiple genre shows versatility, which I expect to correlate positively with a long career.
- number of categories: categories practiced in the first four weeks. We expect effects similar to number of genres.
- n16plus: number of works determined by the site administration to be of adult nature. I have no specific expectations about the effects of this variable.

Variables that reflect peer evaluation are:

- average number of reactions: number of comments received on works written in first four weeks of memberships. I count only those comments received within four weeks of publication of the work. The number of reactions will reflect an author's popularity, perhaps the quality of his writing.
- average rating: the average number of stars given to an author's works. This may again reflect an author being liked sufficiently to want to please him, but it presumably also reflects appreciation for his writing.
- number of ratings: number of works that were rated. One possible hypothesis is that people, rather than rate a story lowly, will refrain from rating it. On that hypothesis, an author that receives a high number of ratings is an author that is being appreciated.
- missing rating: variable expressing whether average rating is missing, because no works were rated. On the same hypothesis as before, this suggests an author who is not appreciated.
- increase in reaction, increase in rating: both for number of reactions and for rating, I computed the correlations with works' rank numbers. A positive correlation here implies an upward trend in number of reactions or rating. All correlations with a significance level below 80% (that is almost all) were set to zero. Presumably, an upward trend in ratings would reflect an increase in quality or an increase in being liked. An increasing number of reactions (very unusual) might be related to an increase in number of 'fans'.

Finally, two variables reflect site evaluation:

- number of selections, number of works of the day: respectively the number of stories that received an exclamation mark and the number of times a work was selected as 'work of the day'. Both presumably reflect quality or at least appreciation on the part of the site administration.

Hypothesis and Test

Candidate independent variables include continued activity on the site after the four initial weeks, the number of works written after the initial weeks, the number of years a user remains active, and the evaluation of later works. Here I discuss continued activity on the site after the four initial weeks and the evaluation of works published after the four initial weeks.

Continuing Activity

The question to be investigated is whether it is possible to predict continued activity on the site after the four initial weeks (i.e. publication of at least one more work) from the data about the first four weeks. As the dataset which we use was created on the basis of all available published works, every author included in the dataset published at least one work. This implies that, among the authors in our dataset, those who did not publish a work in the first four weeks of membership must have done so in the remaining period. This is an effect of the data collection method rather than a true effect of low initial productivity. To account for this situation, authors without publications in the initial four weeks were removed from the dataset. This leaves us with 2315 authors, 1411 of whom (61%) published beyond the initial period. The question to be answered in this case is thus a conditional one, viz.: can we predict, for those authors that publish one or more works in the initial period, whether they will publish at least one work beyond the initial period?

I hypothesize positive effects from the preference for serial works, from user activity (number of works, reactions to others, last active day in the initial period), the variables that would seem to correlate with versatility (number of different genres and categories for the writer's works), variables that indicate positive reception by peers (number of reactions received, average rating, number of ratings, increasing number of reactions) and by the site (number of works of the day). I expect negative effects of a preference for poetry and of recent membership (the self selection effect mentioned above: it is likely that of the older members, only the more active have remained). I test these hypotheses using a logistic regression.

	Stand coef	Sig	
(Intercept)	0.00	0.69458	
date of membership	-1.15	0.00000	***
perc. author replies to comments	0.39	0.00000	***
pref. serial stories	0.22	0.00063	***
no. of works	-1.23	0.00004	***
no. of reactions by author	0.63	0.00132	**
last active day	0.69	0.00000	***
no. of categories	0.32	0.01910	*
avg. no. of reactions received	0.13	0.04028	*
avg. rating received	0.15	0.00930	**
increase in no. of reactions	0.12	0.04611	*
no. of works : last day active	1.51	0.00004	***

Table 2 Standardized coefficients and significance levels for predicting continued activity on site. Significance levels: '***': below 0.1%; '**': below 1%; '*': below 5%; '.': below 10%

Table 2 shows the results of this regression. Variables significant at 5% or better are reported. It appears that the strongest effects come from date of membership, number of works published in the initial period and the interaction between the number of works and the last active day in the initial period. Response by peers (reactions and ratings received), it appears, does have a significant effect, but the effect is small compared to author effects. Perhaps unsurprisingly, the number of reactions given by an author (corresponding, presumably, to an author's willingness to engage with other writers on the site) also correlates with the chance of remaining active on the site.

Evaluation of Later Work

The question to be investigated here is whether it is possible to use behavior and reception in the four initial weeks to predict, after the fact, the evaluation of works published in the remaining period of membership. I did regressions on the four measures of popularity: the average number of reactions received, the average rating, the number of times read and the average time spent in reading. Table 3 reports the standardized coefficients for the four measures.

It is interesting to note how the popularity measures have similar but different determinants. The number of reactions received is only marginally relevant to the rating; conversely, the rating does influence the number of reactions. This suggests that rating is a better measure for appreciation than number of reactions. An author's efforts to engage with others (his reactions to other members' works, his replies to comments) do influence reactions received but influence rating only slightly. Why the percentage of replies to comments should increase the number of times a work is read is not clear to me. A high percentage of poems implies fewer reactions, less positive ratings, and fewer readings. This might imply most members prefer prose. That time spent reading decreases is probably not related to appreciation but to the works' length. People who wrote adult-rated stories are read more but their works are not liked any better for that. The number of categories, which was a predictor for continuing activity, seems unrelated to being appreciated.

	reactions		rating		read		time	
(intercept)	0.0000	**	0.0000	***	0.0000	***	0.0000	.
no. of works	-0.0802	*	0.0012		0.0087		-0.2409	**
avg. no. of reactions received	0.2062	***	0.0057	.	0.0838	***	-0.1080	*
avg. rating received	0.0787	***	0.0147	***	0.0198	.	0.0710	*
No. of ratings received	-0.0302		0.0110	**	-0.0051		0.0254	
date of membership	-0.1682	***	0.0079	**	-0.3479	***	0.1232	***
no. of author's comments to reactions	0.0669	**	0.0004		0.0089		0.0602	
no. of reactions by author	0.0810	***	0.0058	*	0.0227	.	0.0546	
perc. author replies to comments	0.0892	***	0.0029		0.0708	***	-0.0265	
no. of work of the day	0.0487	**	0.0016		0.0049		0.0879	*
no. of selections	0.0107		0.0038		-0.0355	*	0.0696	
pref. stories	-0.0495	.	-0.0087	*	-0.0045		0.0816	.
pref. serial stories	-0.0612	*	-0.0016		-0.0096		0.1853	***
no. of genres	-0.0131		-0.0053	.	0.0127		0.0549	
no. of categories	-0.0010		0.0004		-0.0109		0.0415	
perc. poems	-0.1588	***	-0.0168	**	-0.0608	*	-0.6014	***
no. of 16+ stories	0.0132		0.0024		0.0290	**	0.0211	
increase in rating	0.0339	*	0.0004		0.0066		-0.0164	
increase in no. of reactions	0.0450	**	0.0018		0.0128		-0.0410	
last active day	0.0349	.	0.0056	*	0.0210	.	0.0210	

Table 3 Standardized coefficients and significance levels for predicting later works' evaluation based on data about the first four weeks of activity.

Implications

In many ways, the results raise more questions than they answer. It is clear the last word about the relations between early activity and later success in online writing communities has not been said. The significance of these results is not so much in what they tell us about literary careers on Verhalensite and other online writing communities. The significance is in the fact that by tapping the reservoirs of these sites, literary researchers acquire access to large amounts of data in a field where data used to be very hard to come by. This creates the possibility to, as the SDH 2011 call for papers calls it, ask the formerly unaskable, and it implies a need to rethink what we can expect from literary studies and from literary researchers. Is research along the lines of the above the future for literary studies? Or isn't it even remotely relevant, as it ignores the texts that should be the focus of literary research?

In this discussion, it is important to note that numbers are not the only way of approaching these sites. It should also prove instructive to use text analytical tools for studying which works are successful and which commentators prefer which works. Network analysis tools should help find leaders and followers among the site's participants. That also implies there is a whole new set of tools relevant to the profession of literary studies, and therefore presumably a need for interdisciplinary approaches.

An open question is to what extent the things that we can learn from online writing communities will apply to the regular literary field. Online communities are probably more welcoming and open, less competitive and also perhaps less critical than the literary field as we know it. More research is needed to be able to confirm or reject that hypothesis. In any case, the importance of online writing communities is not limited to their providing a model for our regular literature. As an early example of the communal creativity that the Internet unleashed, online writing communities are also an interesting phenomenon in themselves and in relation to many other internet site types.

Implications for the Research Infrastructure

The source texts of the humanities include not just masterpieces from long ago. Equally important are the many texts written today that may or may not be masterpieces but that can still inform us about literary and cultural processes. These many texts (and, for that matter, photographs, video's, MP3's, etc.) are written, published and reviewed on the web. The web is where today's culture lives and where it should be researched.

For researchers, this is to a large extent a beneficial development. Large amounts of data are available, not tailored for scholarly use, but the data are real, not resulting from contrived situations, and they are nearly free.

These very characteristics also present the researcher with some obstacles. The storage and processing needs are large. Downloading and manipulating large amounts of data require tools and skills that are not widely available among humanities scholars. Similarly, tools for statistical, linguistic or network analysis require an understanding of their underlying models and, importantly, limitations, that humanities scholars are not usually prepared for.

Interdisciplinary research is one obvious response to these obstacles, and it will be to some extent unavoidable, as scholars can't be expected to acquire full expertise in all of the methods potentially helpful in the analysis of these data.

Still, infrastructural facilities supporting the research into online culture could ease significantly the entry of humanist researchers in this field. These facilities should include tools for collecting web pages, for extracting structure out of web pages, for restructuring and enriching data, for viewing and exploring the data and for connecting to other tools. The infrastructure should be cloud-based, as, depending on the type of analysis, the required resources will quickly exceed a personal computer's possibilities.

One of the reasons why a shared infrastructure is especially important is the existence of many similar sites, be they oriented towards text or towards other media. It will be easier to compare and aggregate research results into these multiple sites when the research is based on a shared infrastructure. Another reason is that the existence of this shared infrastructure will help in archiving the considerable amount of data for later researchers. That today's culture lives on the web also implies an urgent need to archive the web on a scale that is far beyond present efforts.

References

Boot, P. (2011). Literary evaluation in online communities of writers and readers [in print]. In Scholarly and Research Communication.

Ekelund, B. G., & Börjesson, M. (2002). The shape of the literary career: An analysis of publishing trajectories. In Poetics, 30 (5-6), 341-364.

Pugh, S. (2005). The democratic genre: Fan fiction in a literary context. Bridgend: Seren.

Van Rees, K., & Vermunt, J. (1996). Event history analysis of authors' reputation: Effects of critics' attention on debutants' careers. In Poetics, 23(5), 317-333.

Verdaasdonk, H. (1985). Empirical sociology of literature as a non-textually oriented form of research. In Poetics, 14(1-2), 173-185.

Verdaasdonk, H. (1987). Effects of acquired readership and reviewers' attention on the sales of new literary works. Poetics, 16(3-4), 237-253.