Contribution, Social networking, and the Request for Adminship process in Wikipedia

Romain Picot-Clémente
UMR CNRS 6285 Lab-STICC

Cécile Bothorel
UMR CNRS 6285 Lab-STICC

Nicolas Jullien
ICI-M@rsouin

Institut Mines Telecom Bretagne
Brest, France
{Romain.PicotClemente, Cecile.Bothorel, Nicolas.Jullien} @telecom-bretagne.eu

1. INTRODUCTION

Epistemic communities are said to be project-oriented communities of experts, evaluated on their contribution in terms of knowledge, where the main criterion for promotion is knowledge production [3]. However, [5], for Wikipedia, [7], for open source, have argued that taking responsibility is an additional step from being a regular contributor, and social interactions with peers may be an additional requirement for being promoted [6].

This work addresses this discussion by looking at the electing process of the administrators (admin) in the English Wikipedia, where exists a quite competitive process of election for the managing position called "administrator", where social connections and knowledge production skills seem to matter. From 2006-01-01 to 2007-10-01, which is our period of study, there were 1,617 RfA, with a 49.2% rate of success).

Burke and Kraut [2] proposed a model to predict RfA results[1], mainly based on counting attributes for modelling the candidate's activity, according to the criteria put forward by the Guide to RfAs[2]. Their model's accuracy reached 75.6%. However, they did not measure the respective influence of the edits and of the social interaction on RfA results, as their measure of social influence is based on the discussion in the article pages, which can be considered as production-related activities more than social interaction. Neither did they separate social networking with administrators from social networking with everyone, whereas an administrator (or a bureaucrat) may be more influential than an unknown user on an RfA result [4]. This work aims at addressing these limitations. It provides more accurate model than [2] (78% vs 75.6%), with more simple variables, easier to extract from the raw data. It shows that beyond a required minimal number of edits, social interactions with people and with peers make the difference between success and failure in the RfA.

[1]Using the Stanford Large Network Dataset Collection, `https://snap.stanford.edu/data/##wikipedia`, as we did, to be comparable.

[2]`http://en.wikipedia.org/wiki/Wikipedia:GRFA`

OpenSym August 19-21, 2015, San Francisco, USA
2015 ACM 978-1-4503-3666-6/15/08
DOI: http://dx.doi.org/10.1145/2788993.2806211 .

2. MODEL

2.1 The variables

Revision Activities.

We extracted the number of revisions/editions made (variable: *Revision*), the number of distinct pages edited (*Pages*) and the number of distinct categories edited (*Categories*), and the repartition of the revisions ($Revision_{repartition}$) in order to take into account both the volume and the variety of the revisions (Gini coefficient on the number of revisions per pages).

We assumed that the talks on the articles' discussion pages were related to the revision activities: the number of distinct pages the candidates talked on (*TalkPages*), the total number of talks on the articles' discussion pages (*PageTalks*), and the Gini Coefficient fo their repartition ($PageTalks_{repartition}$).

Social Activities.

We focused on the conversations on the users' pages to assess the impact of non-knowledge-production interactions. We created three weighted and oriented graphs: a general one, named *userSN* (nodes: the considered candidates and all the wikipedians); *adminSN* (nodes: the considered candidates and the (already) admins); *burSN* (nodes: the considered candidates and the (already) bureaucrats).

For each graph, we computed the attributes that described the characteristics of each node "candidate". They have the same name, with as suffix the name of the related graph: the degree of the node (*Degree*), without taking into account the orientation of edges; the number of distinct users (resp. admins, bureaucrats) to whom the candidate posted a message on their user page (*outDegree*), the number of distinct users (resp. admins, bureaucrats) that posted a message on the candidate's page (*inDegree*); the total number of messages posted (*outTalksNumber*) and received (*inTalksNumber*) by the candidate. Then, we computed multiple centrality measures on the graphs: *Closeness*, *PageRank*, and *Betweenness*. Finally, we computed the Gini coefficients for both the number of messages posted by the candidates ($outTalks_{repartition}$), and received by them ($inTalks_{repartition}$).

2.2 The models

We used the random forest algorithm [1] to predict the election, with a learning population (70% of our sample). Since we want to understand the contribution of the social attributes in the RfA result, we first create two predictive models, one based on the revision attributes (Model 1) and one based on the social attributes (Model 2). Then, we consider a model using all the attributes (Model 3, not presented here), and then Model 3's most relevant attributes (Model 4, which is as good as Model 3, but with less variables and a smaller variance in accuracy).

3. RESULTS

According to the results presented in Table 1, social and revision attributes seem to be complementary for predicting the RfA results.

	Global Accuracy	Confusion Matrix			
			0	1	Accuracy
Model 1 (Revisions)	74%	0	169	77.5	68.6%
		1	48	190	79.8%
Model 2 (Social)	74%	0	168	80	67.7%
		1	46.5	192	80.5%
Model 4 (Rev. + soc.)	78%	0	177	70	70.6%
		1	37	200	84.4%

Table 1: Accuracy and Confusion matrix.

The random forest method computes the average decrease of accuracy for each tree in the forest when a given attribute is not used. According to this metric, the most important attributes are *Revisions*, *TalkPages*, $outDegree_{userSN}$. Table 2 details those attributes for predicting successful and unsuccessful promotions.

For each of the main 9 attributes, but the PageRank, there are significant behavioural differences between the promoted and the non-promoted candidates: the interdecile range of the probability density for the promoted candidates is smaller than the one for non-promoted candidates (the curve is flatter). This suggests that the promoted candidates behave more similarly than the non-ones, explaining why the models predict better promotion than non-promotion. The successful candidates are also more active than the unsuccessful ones (for example a candidate with a greater $outDegree_{userSN}$ is more likely to be elected).

The estimated probability of being elected according to the number of revisions (*Revisions*) is less than 50% beyond about 2500 revisions and is about 70% beyond 6000 revisions. There is a similar behavior pattern (probability of 50% beyond a threshold (T) and extreme values with not enough case studies) for the following attributes: *Categories* ($T\approx1700$), *TalkPages* ($T\approx130$), $outDegree_{userSN}$ ($T\approx450$), $outDegree_{adminSN}$ ($T\approx17$), $inDegree_{userSN}$ ($T\approx140$).

Successful promotions	Unsuccessful promotions
outDegree_userSN	Revisions
Revisions	TalkPages
TalkPages	inDegree_userSN
outDegree_adminSN	Categories
Categories	outDegree_adminSN

Table 2: The most important attributes to predict successful and unsuccessful promotions

4. DISCUSSION

Our results are consistent with the Guide to RfA, previous results and the theories on epistemic communities. Regarding the guide, we provide much more precise figures of how many contributions and interactions are needed to have a high probability for being elected. We show that there are quite narrow windows in terms of number of contributions and discussions, in which the chances of being elected are maximized. We dramatically simplified the measures proposed by [2] regarding the edit activities, with a better evaluation of the chances to be elected (78% of good prediction vs 75.6%), while keeping the number of explanatory variables reasonably low, and easy to extract from the raw data.

As supposed for an epistemic community, the contribution in knowledge (*Revision*) is the first criterion to be considered as a good candidate, and a shortage of contribution is often synonymous to failure. But, once the candidates have proven their competence (production of knowledge) and their willingness to do the job (interacting with people), knowing and being known by these core members, who are the future peers, makes the difference.

There are obvious limitations to our work which will be addressed in future work: we studied one project only and with not recent data. Second, if our model is good at forecasting the elections (more than 80% of accuracy), it is not as good for the non-elections (around 70%). Dropping the extreme cases (people who are not elected because they talked too much, maybe because they fought too much) may improve the prediction.

References

[1] L. Breiman. Random forests. *Machine learning*, 45(1): 5–32, 2001.

[2] M. Burke and R. Kraut. Mopping up: Modeling wikipedia promotion decisions. In *Proceedings of CSCW 2008*, pages 27–36. ACM, 2008.

[3] P. Cohendet, F. Créplet, and O. Dupouet. Interactions between epistemic communities and communities of practice as a mechanism of creation and diffusion of knowledge. In J.-B. Zimmermann and A. Kirman, editors, *Interaction and Market Structure*. Springer, Londres, 2001.

[4] J. B. Lee, G. Cabunducan, F. G. Cabarle, R. Castillo, and J. A. Malinao. Uncovering the social dynamics of online elections. *Journal of Universal Computer Science*, 18(4):487–505, 2012.

[5] C. Pentzold. Imagining the Wikipedia community: What do Wikipedia authors mean when they write about their "community"? *New Media & Society*, 13(5):704–721, August 2011.

[6] F. Rullani and S. Haefliger. The periphery on stage: The intra-organizational dynamics in online communities of creation. *Research Policy*, 42(4):941–953, 2013.

[7] G. von Krogh, S. Haefliger, S. Spaeth, and M. W. Wallin. Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development. *MIS Quarterly*, 36(2):649–676, 2012.