Using WikiProjects to Measure the Health of Wikipedia

Ramine Tinati, Markus Luczak-Roesch, Nigel Shadbolt, Wendy Hall
Web and Internet Science
University of Southampton
{r.tinati|mlr1m12|nrs|wh}@ecs.soton.ac.uk

ABSTRACT

In this paper we examine WikiProjects, an emergent, community-driven feature of Wikipedia. We analysed 3.2 million Wikipedia articles associated with 618 active Wikipedia projects. The dataset contained the logs of over 115 million article revisions and 15 million talk entries both representing the activity of 15 million unique Wikipedians altogether. Our analysis revealed that per WikiProject, the number of article and talk contributions are increasing, as are the number of new Wikipedians contributing to individual WikiProjects. Based on these findings we consider how studying Wikipedia from a sub-community level may provide a means to measure Wikipedia activity.

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous

Keywords

Wikipedia; WikiProjects; Social Machines

1. INTRODUCTION

Wikipedia, the global online encyclopedia, has gained both academic and pubic attention given its role as a highly successful, collaboratively-produced knowledge resource. However, despite its widely acknowledged value, questions have been raised about the state and health of Wikipedia, with the most recent findings suggesting a relatively small and decreasing number of contributors [14], and the overall generation of new contributions may be in a state of decline [5].

A decade after its launch, Wikipedia has managed to retain much of its technical core. Additionally, the system has witnessed several community-led reformations of how the technical system is used to support and coordinate the work of the volunteer editors. One interesting feature are the more than 2,000 WikiProjects. Representing an attempt to improve the breadth and depth of Wikipedia articles in a multitude of specific domains, about one million Wikipedians are contributing to millions of Wikipedia pages by coordinating work via WikiProjects.

In this paper we examine if the emergent WikiProjects may provide an alternative perspective on understanding the general liveliness of Wikipedia activity. Our analysis is the first attempt to investigate the current state of health of Wikipedia using WikiProjects as a proxy of activity. We show that this is a promising new way of looking at Wikipedia, suggesting that a kind of knowledge saturation of the actual article contents upgrades the importance of coordination mechanisms such as WikiProjects for both, experiences editors as well as newcomers.

2. BACKGROUND AND RELATED WORK

Wikipedia studies are varied, from examining the structure articles and contributors [13], to the motivations, participation and demographics of contributors, to editing conflicts and disputes [7], to barriers to adoption [6]. Efforts have been made to understand the collaborative process of article creation [8], and devising methods to measure the quality of an article [1]. Studies have examined the dialogue and coordination of contributors in these environments [3], and the role of discussion in the production of an article [4, 10], as well as facilitating communication with other editors [12].

WikiProjects is a community-driven feature that enables contributors to work together to create and improve articles, usually related to a specific area of interest or domain. There are no formal guidelines on how a WikiProject should be run, and there are no privileges for those contributing to a WikiProject. The articles which a WikiProject is contributing to is also part of the main corpus of Wikipedia articles, and can be part of more than one WikiProject. A WikiProject uses a specific template for its home (root) page. This template – developed by Wikipedians – resembles a typical Wikipedia article structure, but contains specific content related to a project, including, contributors, articles associated with, and the 'project goal'. Initial studies of WikiProjects have examined team coordination [11], and motivation to contribute [2].

3. EXPERIMENT SETUP

This study analysed 1.6 million English Wikipedia article pages and their corresponding Talk page. We harvested over 3.2 million unique wiki pages, which are associated with 618 active WikiProjects. Although there are 2000 projects, only active WikiProjects marked by Wikipedia were harvested. In addition to the harvest of all Wikipedia articles and Talk pages associated with the 618 WikiProjects, we also harvested all the 'root' pages, which are the homepage of the WikiProject.

We harvested all Wikipedia pages related to all 618 active WikiProjects. For each Wikipedia page, we harvested the log of all edits. Each revision contained details of the user, time, and the change made. If the user was not registered, then the user's IP address is recorded, and 'anon' is appended to the entry. We also extracted all corresponding 'Talk' pages if available (not all articles have a Talk page with entries). For each Talk entry, the user, the timestamp, and the comment made is stored. We then compute the growth of a project by constructing the timeseries of edits, talk entries, and new users since the earliest project page, relative to the first entry of the root project page.

Copyright is held by the author/owner(s).
WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2745937 .

Measure	All Project Pages	Root Pages
Avg # of articles	1,718	1
Avg # of Talk Pages	1,700	1
Avg # of article Entries	177,533	453
Avg # of Talk Entries	25,028	234
Avg # of Wikipedians	43,036	117
Avg # of article editors	40,294	117
Avg # of Talkers	4,817	62
Avg # of Anon. editors	1,308	51
Avg # of Anon. Talkers	2,324	46
Avg # of Crossover Wikipedians (%)	1,868	12
Avg % are Crossovers Wikipedians	4.5	13

Table 1: WikiProject Statistics. November 2014. Root pages represent the activity on the core WikiProject page (homepage)

Feature	Avg growth function	Std Dv.
All article pages in Project	2.58x - 2.8	0.12
All Talk pages in Project	2.92x - 3.31	0.25
Root article pages in Project	3.68x - 4.32	0.74
Root Talk in Project	3.86x - 4.76	0.60

Table 2: Regression Analysis of Editing and Talk Activity Growth for All and Root Pages

4. RESULTS AND FINDINGS

We compared the 618 WikiProjects with respect to the set of articles, Talk contributions, and users associated with a given WikiProject. As shown in Table 1, we extracted users who have made entries on the 'root' WikiProject page and also for the total set of pages related to a project.

Comparing the WikiProjects, we found a normal distribution with regards to the number of Wikipedians, articles and Talk pages, and found a positive correlation between the number of Talk entries (for a Wikipedia in the root page) and the number of articles edited. Unlike the main corpus of Wikipedia article [15], the number of edit and talk contributions per user for root pages of a WikiProject, were more distributed, which may be due to type of content that these pages contained.

We computed the growth of each WikiProject in terms of the number of contributions, pages, and newly joined Wikipedians. Results in Table 2 show that WikiProjects follow a linear growth function (article and Talk pages). The S.D. indicates that WikiProjects grow at a similar rate, which is also true for the set of root pages. In comparison to Suh et al. [14] and Halfaker et al. [5], our findings suggest that based on the WikiProject activity, Wikipedia is not in decline, but still enjoying growth with new users, edits, and discussion activity. Akin to other complex online communities [9], using traditional methods to measure community and system health may not reflect their true state; instead, we need to develop novel techniques to examine the evolving social machinery of Wikipedia.

5. CONCLUSION

Our study has shown that by using WikiProjects as a proxy for examining contribution and discussion activity, Wikipedia can be described as an active social machine with many thriving domain specific, sub-communities. We have noted growth across the WikiProjects various domains of interest, and see how talk has become significant as a project develops.

Although further analysis is required to understand the context and relationship between article production, discussion, and the role of WikiProjects, our initial findings suggest that WikiProjects are supporting the creation and improvement of Wikipedia articles, and that Talk is an important feature that facilities the communication and interaction within the WikiProject sub-communities.

6. REFERENCES

[1] De la Calzada, G., and Dekhtyar, A. On measuring the quality of wikipedia articles. In *Proceedings of the 4th Workshop on Information Credibility*, WICOW '10, ACM (New York, NY, USA, 2010), 11–18.

[2] Farič, N., and Potts, W. H. Motivations for contributing to health-related articles on wikipedia: An interview study. *J Med Internet Res 16*, 12 (Dec 2014), e260.

[3] Ferschke, O., Gurevych, I., and Chebotar, Y. Behind the article: Recognizing dialog acts in wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, Association for Computational Linguistics (Stroudsburg, PA, USA, 2012), 777–786.

[4] Forte, A., Kittur, N., Larco, V., Zhu, H., Bruckman, A., and Kraut, R. E. Coordination and beyond: Social functions of groups in open content production. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, ACM (New York, NY, USA, 2012), 417–426.

[5] Halfaker, A., Geiger, R. S., Morgan, J., and Riedl, J. The rise and decline of an open collaboration system: How wikipedia's reaction to sudden popularity is causing its decline. *American Behavioral Scientist 57*, 5 (May 2013), 664–688.

[6] Hautasaari, A., and Ishida, T. Analysis of discussion contributions in translated wikipedia articles. 57–66.

[7] Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. He says, she says: Conflict and coordination in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, ACM (New York, NY, USA, 2007), 453–462.

[8] Liu, J., and Ram, S. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Trans. Manage. Inf. Syst. 2*, 2 (July 2011), 11:1–11:23.

[9] Luczak-Rösch, M., Tinati, R., Simperl, E., Kleek, M. V., Shadbolt, N., and Simpson, R. Why won't aliens talk to us? content and community dynamics in online citizen science. In *Eighth International AAAI Conference on Weblogs and Social Media* (2014).

[10] Morgan, J. T., Gilbert, M., McDonald, D. W., and Zachry, M. Project talk: Coordination work and group membership in wikiprojects. In *Proceedings of the 9th International Symposium on Open Collaboration*, WikiSym '13, ACM (New York, NY, USA, 2013), 3:1–3:10.

[11] Morgan, J. T., Gilbert, M., Zachry, M., and McDonald, D. A content analysis of wikiproject discussions: Toward a typology of coordination language used by virtual teams. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion*, CSCW '13, ACM (New York, NY, USA, 2013), 231–234.

[12] Oxley, M., Morgan, J. T., Zachry, M., and Hutchinson, B. "what i know is...": Establishing credibility on wikipedia talk pages. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, WikiSym '10, ACM (New York, NY, USA, 2010), 26:1–26:2.

[13] Stuckman, J., and Purtilo, J. Measuring the wikisphere. *Proceedings of the 5th International Symposium on Wikis and Open Collaboration* (2009).

[14] Suh, B., Convertino, G., Chi, E. H., and Pirolli, P. The singularity is not near: Slowing growth of wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, WikiSym '09, ACM (New York, NY, USA, 2009), 8:1–8:10.

[15] Voss, J. Measuring wikipedia. *International Conference of the International Society for Scientometrics and Informetrics: 10th, Stockholm (Sweden)* (2005), 24–28.