Measuring the Quality of Edits to Wikipedia

Susan Biancani
Stanford University
520 Galvez Mall
Stanford, CA 94305
biancani@stanford.edu

ABSTRACT

Wikipedia is unique among reference works both in its scale and in the openness of its editing interface. The question of how it can achieve and maintain high-quality encyclopedic articles is an area of active research. In order to address this question, researchers need to build consensus around a sensible metric to assess the quality of contributions to articles. This measure must not only reflect an intuitive concept of "quality," but must also be scalable and run efficiently. Building on prior work in this area, this paper uses human raters through Amazon Mechanical Turk to validate an efficient, automated quality metric.

Categories and Subject Descriptors

H.5.3 [Group and Organization Interfaces]: Computer-supported cooperative work, web-based interaction

H.1.2 [Models and Principles]: User/Machine Systems

General Terms

Measurement, Human Factors.

Keywords

Wikipedia, Peer, Peer Review, WikiWork, Experience, Ownership, Quality, Collaboration

1. INTRODUCTION

Wikipedia's value is not disputed. It is a widely used and trusted reference, and is among the world's most popular websites. JStor's Data for Research site lists 390 papers that cite Wikipedia in their references, many of which are not about Wikipedia: it appears scholars have begun to cite Wikipedia, often as a source providing the reader with reliable background information about a given topic.

While its success is not in dispute, it is incredibly puzzling. How can a site that allows anyone to write anything, and that makes all contributions live immediately—without first sending them through some approval process—possibly end up full of useful, informative content, and not nonsense and vandalism?

OpenSym '14 , Aug 27-29 2014, Berlin, Germany
ACM 978-1-4503-3016-9/14/08.
http://dx.doi.org/10.1145/2641580.2641621

As a peer-production community, Wikipedia bears less resemblance to traditional productive organizations than to an informal social grouping. It relies, like many social groups, on reinforcement of social norms by community members. These norms are clearly articulated on the site in the form of collectively produced policy pages, such as that describing the three "Core Content Policies" that define the characteristics of good contributions. These pages are open to editing by all and reflect the broad consensus of the editor community. I hypothesize that the expression and enforcement of these norms, through messages sent between users on Wikipedia, shape the quality of future edits, and contribute to the maintenance of the encyclopedia's value. Each Wikipedia editor has a user page, on which others can leave comments and notes. Frequently, editors use this space to provide feedback to the editor who owns the page. I argue that these messages often convey social norms shared by the broad community of Wikipedia editors, and predict that these messages can affect the quality of the recipient's future edits.

In order to pursue this line of theory, I need to identify a reliable, automated measure of the quality of editor contributions to Wikipedia. Such a measure would be applicable to a wide range of research questions pertaining to Wikipedia, to its editor community, and to its value as a reference. In this paper, I present work in progress to validate an existing measure of the quality of revisions to Wikipedia.

2. RELATED WORK

Scholars have been keenly interested in the question of how Wikipedia maintains the quality of its articles almost since the site's inception. In particular, researchers are interested to identify social mechanisms related to an article's development that predict its quality. Kittur and Kraut [5] ask how different modes of coordinating editing tasks affected the quality of articles. Liu and Ram [6] investigate the effects of different collaboration patterns among editors assuming a variety of roles. [9] examine the effect of intensive periods of cooperation on articles. In all of these studies, the quality of the resulting article is treated as the dependent variable. Quality in these studies is operationalized by recording whether articles have attained one of several designations applied by the Wikipedia editor community, such as "featured article" status, "good article" status, or one of several other grades.

Reliance on these designations can have a limiting effect on research. Because these grades rely on human judgment achieved through consensus, a minority of articles are ever graded, limiting the study population and potentially making it difficult to generalize the findings. Those that do receive grades are infrequently updated, making it difficult to study processes of improvement or degradation. Finally, the grades reflect the quality of the article as a whole, and cannot be used to study the effect of particular contributions, or to assess the efficacy of individual editors.

To address these concerns, scholars have sought to develop automatic measures of article quality and of edit quality. One of the earliest of these was word count, which [1] demonstrated correlated with the likelihood of receiving featured article status. While word count may work relatively well across a population of articles, it is an overly simple metric with which to assess the quality of any particular article. [10] used features of the lifecycle of articles, and [2] used a combination of length, structural features, and stylistic features. Both of these approaches improved on word count as a measure of the overall quality of an article without offering a clear way to assess the quality of contributions or of editors.

The most promising efforts to evaluate the quality of contributions rely on measures of word persistence. The quality of editors can then be calculated by aggregating the quality scores over the collection of all the editor's contributions. Halfaker et al. [3] devised Persistent Word Revisions (PWR). For a given contribution to an article (i.e. a single act of revision to the article), they follow the words in the contribution forward over later edits, counting the number of subsequent revisions through which each word survives. Priedhorsky et al. [7] use a similar measure, counting the length of real-time the contribution survives. Because Priedhorsky et al.'s approach does not normalize for differing rates of edit activity across articles, Halfaker et al.'s metric is preferred.

Finally, [8] further refine this approach, by iteratively assessing both the quality of contributions and the quality of contributors, each informing the other until convergence is reached. While this method is compelling, at present it is not scalable to the full size of English-language Wikipedia. In contrast, PWR scales well, and the code to generate scores is freely available from the first author's code repository.

In introducing PWR, the authors presented the results from limited validation of their measure, using the same article status categories as Kittur and Kraut [5]. Specifically, they created a sample of articles that had improved over time, as indicated by human-assigned quality grades. They then asked whether articles that had improved were more likely to have been edited in the interim by editors with high overall word-persistence scores. While their results supported their hypothesis, this approach offers minimal validation of PWR as a measure. Multiple editors had contributed to each article over the period between grade assessments. Their contributions were aggregated, and variation among these editors washed out. Thus, confirmation of PWR as a metric is indirect. Moreover, the population of articles that received multiple grades over the study period was small and non-random, and findings may not generalize well to the encyclopedia as a whole.

Although PWR has not been rigorously validated, it is nonetheless a compelling measure of edit quality: it is conceptually straightforward, efficient to calculate, and computationally tractable. As such, it has the potential to be profoundly useful to Wikipedia researchers. In this study, I offer the results of my effort to rigorously validate this measure.

3. METHODS & PRELIMINARY RESULTS

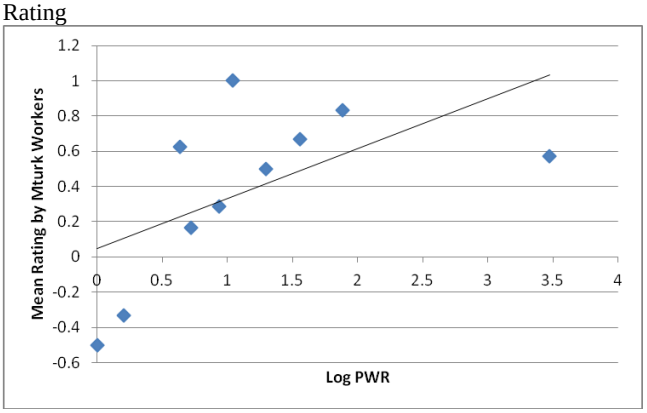
Using code freely available from Aaron Halfaker's repository, I generated PWR scores for a random sample of edits to Wikipedia. I then ranked edits into deciles on the basis of their PWR scores, and sampled one edit from each decile. (In this paper I present the results of a pilot study, but at the OpenSym conference I will present the results of a much larger study, using the same approach but a much larger number of edits.)

Workers from Amazon Mechanical Turk rated each edit in the sample. Workers were given a link to the Wikipedia page showing the comparison between revisions, comparing the article just before the edit in question to the same article just after. Workers were asked to visit the link, consider the revision, and then respond to the following forced-choice question:

> What effect did the revision you considered have on the quality of the Wikipedia article?
>
> It **improved** the article's quality.
>
> It **worsened** the article's quality.
>
> It had **little or no effect** on the article's quality.
>
> I'm not sure.

I scored responses as follows: "improved" = 1, "worsened" = -1, "no effect"/"not sure" = 0. I then computed the mean score for each revision tested. Figure 1 shows a scatter plot of the mean score by MTurk respondents versus the log PWR score, with trendline added.

Figure 1. Scatterplot of Log PWR versus mean Amazon Turk Rating



These results are based on a small pool of 63 total ratings of ten revisions. While this dataset is certainly too small to be conclusive, the early results are promising.

4. FUTURE WORK

Having run a first pilot of my study, I plan to refine its design and presentation in several ways. I plan to remove the "no effect" rating, leaving only "improve", "worsen", and "not sure". I will also design a brief training for workers to complete before starting to rate articles. This training will highlight some of the important features of the Wikipedia layout of pages that compare between revisions. For example, that added text appears in bold, and highlighted in blue, while deleted text is highlighted in yellow. Finally, I will create a pool of test cases (revisions that should be straightforward to understand) for workers to rate before allowing them to rate the full sample. After I have tested each of these components on a small sample of revisions, I will repeat the study on a much larger sample of revisions. This is the work I intend to present at OpenSym 2014.

5. ACKNOWLEDGMENTS

Many thanks to Aaron Halfaker for generously sharing the code to generate PWR scores, and for extensive advice and trouble-shooting.

This work was supported by a Dissertation Support Grant from the Stanford Graduate School of Education.

6. REFERENCES

[1] Blumenstock, J. E. (2008). Size matters: word count as a measure of quality on wikipedia. *Proceedings of the 17th international conference on World Wide Web*, 1095–1096. Retrieved from http://dl.acm.org/citation.cfm?id=1367673

[2] Dalip, D. H., Gonçalves, M. A., Cristo, M., & Calado, P. (2009). Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries* (pp. 295–304). Retrieved from http://dl.acm.org/citation.cfm?id=1555449

[3] Halfaker, A., Kittur, A., Kraut, R., & Riedl, J. (2009). A Jury of Your Peers : Quality , Experience and Ownership in Wikipedia. *Proceedings of the International Symposium on Wikis and Open Collaboration*.

[4] Hu, M., Lim, E.-P., Sun, A., Lauw, H. W., & Vuong, B.-Q. (2007). Measuring article quality in wikipedia: models and evaluation. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 243–252). Retrieved from http://dl.acm.org/citation.cfm?id=1321476

[5] Kittur, A., & Kraut, R. E. (2008). Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. *Proceedings Of Computer-Supported Cooperative Work*, 37–46.

[6] Liu, J., & Ram, S. (2009). Who Does What: Collaboration Patterns in the Wikipedia and Their Impact. *19th Workshop on Information Technologies and Systems* (pp. 175–180).

[7] Priedhorsky, R., Chen, J., Lam, S. (Tony) K., Panciera, K., Terveen, L., & Riedl, J. (2007). Creating, destroying, and restoring value in wikipedia. *Proceedings of the 2007 international ACM conference on Conference on supporting group work - GROUP '07* (p. 259). New York, New York, USA: ACM Press. doi:10.1145/1316624.1316663

[8] Suzuki, Y., & Yoshikawa, M. (2012). Mutual evaluation of editors and texts for assessing quality of Wikipedia articles. *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration - WikiSym '12*. New York, New York, USA: ACM Press. doi:10.1145/2462932.2462956

[9] Wilkinson, D., & Huberman, B. (2007). Cooperation and quality in wikipedia. *Proceedings of the 2007 international symposium on Wikis - WikiSym '07* (pp. 157–164). New York, New York, USA: ACM Press. doi:10.1145/1296951.1296968

[10] Wöhner, T., & Peters, R. (2009). Assessing the quality of Wikipedia articles with lifecycle based metrics. *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. Retrieved from http://dl.acm.org/citation.cfm?id=1641333