Evaluating ORES:

A Machine-learning-based Article-quality Classifier for Wikipedia

Zhihan Yang, Carleton College

Background

Launched in 2001, Wikipedia has become the most well-known online encyclopedia. According to Alexa Internet, Wikipedia is the 5th most visited website on the global Internet by the end 2016. Correspondingly, Wikipedia enjoys a large contributor base and a huge number of articles. As reported by WikiStats, Wikipedia has 3.6 million articles available in English and 7.4 million articles available in other languages by June 2011.

The rise of Wikipedia has led to the decline or demise of its ancestors, such as Nupedia and *Encyclopædia Britannica*. Arguably, what set Wikipedia apart from its ancestors is that, instead of using an expert-centered production model, Wikipedia took advantage of the peer-production model by setting a low barrier of entry, i.e., anyone with Internet access and a device with convenient web-browsing functionality can access and contribute to Wikipedia easily.

Wikipedia's openness to contributions is not without its costs. Since most contributors are not experts and receive no external incentives, the quality of contributions is highly varied. The assortment of contributions make it very difficult for a relatively small number of administrators to evaluate the quality of and decide what to do with each and every contribution. Fortunately, a variety of algorithmic tools have been created to ameliorate this problem. These tools are of paramount importance because they allows fast identification and correction of low-quality contributions. However, little literature has analyzed the effectiveness of such tools.

Motivation and Research Question

Moderation on Wikipedia encompasses two categories of tasks and different algorithmic tools has been developed for each of the two categories. One category of tools can be thought of as "bounty hunters". They detect the presence of specific words that hint at potential vandalisms. The metrics used by these tools are explicitly defined by the lists of unwanted words given by human administrators. Their performances can evaluated in a straightforward manner by, for example, monitoring the rate at which vandalisms are spotted by non-administrators and the average time elapsed between a vandalistic contribution and its reversion.

Another category of tools can be best regarded as "intelligent assistants" for human administrators. They are machine-learning-based tools capable of learning decision metrics from human-labelled data and are deployed to perform cognitively demanding tasks, such as evaluating the quality of articles. These tools rely on more sophisticated and less interpretable methods, and can be much less accurate when processing data that are very different from training data. A 2009 study of physicians showed that Wikipedia is used in 26% of patient cases and by 70% of physicians for clinical purposes (Hughes et al, 2009). Therefore, having a clear sense of how these less interpretable tools are performing is crucial because articles are applied to high stake decision-making scenarios such as diagnosis of illnesses.

One such machine-learning-based tool is the Objective Revision Evaluation Service (ORES), a web service created and maintained by Wikimedia Scoring Platform team for quality evaluation of edits and articles. The goal of this study is to investigate the relationship between the number of external links or citations, a widely used metric in academia for measuring validity of articles, and the quality classification made by ORES. This study also explores the

relationship between quality classification and number of pageviews, an intuitive measure of impact. If ORES is an effective tool for article-quality evaluation, the quality classification received by an article should strongly correspond to its number of external links and number of pageviews. By comparing results with this hypothesis, this study reveals how well reality aligns with the ideal and how concerned we should be. Based on conclusions obtained from statistical analysis, this study proposes concrete suggestions on how ORES should be used by administrators on Wikipedia.

Data Collection

The number of external links and the total number of pageviews (over 60 days, from March 23rd 2019 to May 22nd 2019) were collected for 10367 articles under the category Medicine. Data were collected using the MediaWiki API. Among all articles under Medicine, I excluded ones under subcategories including Medical activism, Medical associations, Medical culture, Medical education, Health insurance, Works about medicine, Medicine stubs and their subcategories because they are not pertinent to clinical diagnosis. The code used for data collection was elucidated in my previous work (Yang, 2019).

In addition to number of external links and total number of pageviews, the quality classification made by ORES was also collected for each one of the 10367 articles. Data were collected using the third version of the ORES scoring interface API. This API only receives revision IDs as inputs. A revision ID is a unique tag given to a past snapshot or version of an article. Revision IDs of the most recent versions of the 10367 articles were first collected using the MediaWiki API and were then inputted into the ORES API for quality evaluation. For a

revision ID, ORES API processes its corresponding article and returns a dictionary that maps from possible quality categories (from the best to the worst: FA, FL, A, GA, B, C, Start and Stub) to probabilities; the quality category assigned with the highest probability was considered the ultimate quality classification for that article.

Exploratory Data Analysis

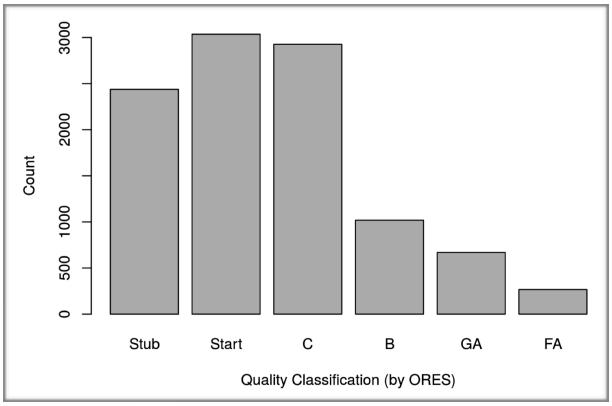

Figure 1. Histogram of quality classifications made by ORES. Quality categories are arranged from the worst (Stub) to the best (FA).

The histogram of *quality classifications made by ORES* (Fig. 1) shows that the distribution of quality classification is quite right-skewed, indicating that the majority of articles are clustered at lower qualities ("C" and "Start") and higher quality classifications ("B", "GA" and "FA") are assigned to relatively fewer articles. It should be noted that quality categories FL (between FA and GA) and A (between GA and B) are not assigned to any articles.

The 5-number summary of *the number of pageviews* are 3 (minimum), 402 (25th quantile), 1674 (median), 7345 (75th quantile) and 787314 (maximum). The mean of the number of page views is 10137. Since the mean is significantly greater than the median, the distribution of page views is extremely right-skewed.

The 5-number summary of *the number of external links* are 1 (minimum), 6 (25th quantile), 13 (median), 30 (75th quantile) and 501 (maximum). The mean of the number of

external links is 27. Since the mean is noticeably greater than the median, the distribution of external links is significantly right-skewed.

Further Analysis

One goal of this study is to explore the relationship between number of external links, a metric for determining credibility of articles, and quality classification made by ORES. Number of external links acts as the baseline for determining the quality of evaluations made by ORES. Since both measures are intended to reflect the credibility of articles, articles with higher number of external links should be, on average, rated higher by ORES.

This study also looks into the relationship between number of pageviews, an indicator of impact, and the quality classification made by ORES. Yang (2019) showed that log-transformed number of external links is able to explain about 23% of the variance of log-transformed number of pageviews. It would be interesting to see whether quality classifications allows for more accurate predictions than log-transformed number of external links.

Relationship between number of external links (NEL) and quality classifications made by ORES (QC). The box-plot of NEL by QC (Fig. 4) shows that the distribution of NEL for each quality category is right-skewed, which hints at log-transforming NEL. This has an immediate advantage. As shown in Fig. 5, distributions of log-transformed NEL have similar variances and are symmetric, which satisfies the assumption of and thus allows for modeling using a linear regression model. Fig. 5 reveals that the median of log-transformed NEL increases as quality classification improves, except at quality category “GA”. To further illustrate this phenomenon, Fig. 7 shows increments of medians and means of log-transformed number of

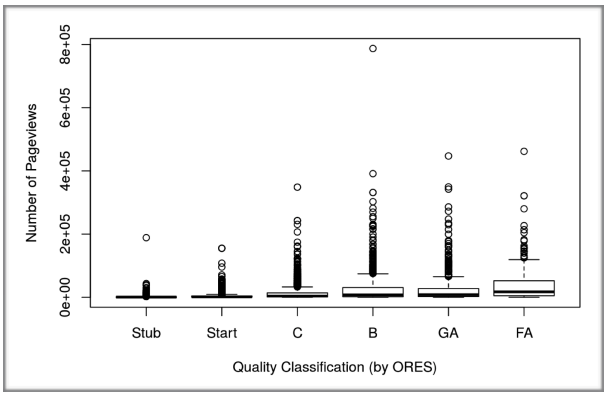

Figure 2. Box-plot of number of pageviews by quality classifications. The density of data points is clearly higher towards zero, which hints at log-transforming number of pageviews.

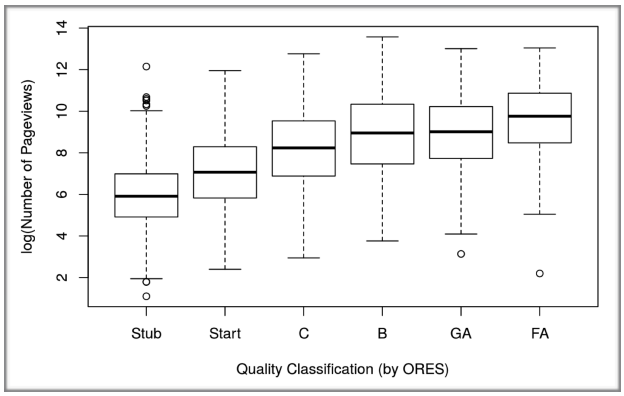

Figure 3. Box-plot of log-transformed number of pageviews by quality classifications. The value of the median increases as the quality classification improves.

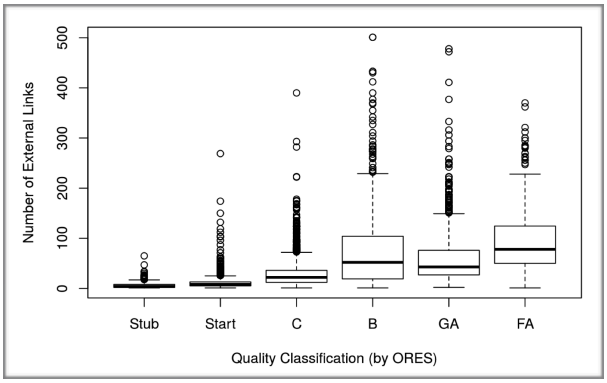

Figure 4. Box-plot of number of external links by quality classifications. The density of data points is clearly higher towards zero, which hints at log-transforming number of external links.

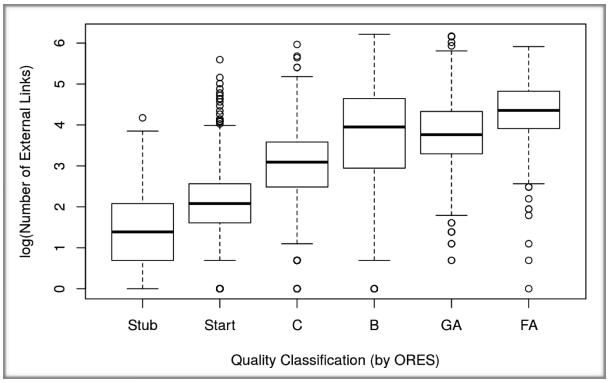

Figure 5. Box-plot of log-transformed number of external links by quality classifications. The value of the median increases as the quality classification improves, except at "GA".

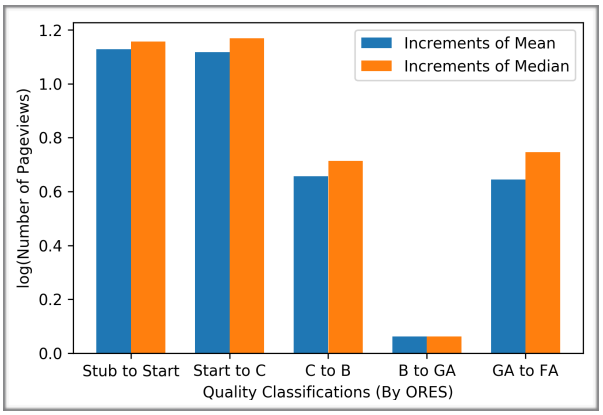

Figure 6. Side-by-side box-plot of increments of medians and means of log-transformed number of pageviews between neighboring quality classifications. Note that both increments are smallest from "B" to "GA".

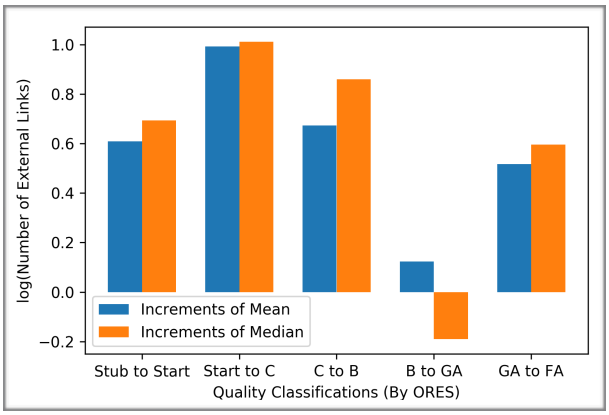

Figure 7. Side-by-side box-plot of increments of medians and means of log-transformed number of external links between neighboring quality classifications. Note that the increment of mean is smallest from "B" to "GA" and that the increment of median is negative from "B" to "GA".

external links between neighboring quality categories. The increment of mean from “B” to “GA” is relatively small compared to other increments of mean; the increment of median from “B” to “GA” is negative while all other increments of median are positive.

Regression with QC as the explanatory variable and log-transformed NEL as the response variable shows that QC can explain 47.2% variance in log-transformed NEL. This indicates that the QC of an article has a high predictive power for its number of external links and vice versa. The results are shown in Table 1. Since this is not an inference problem and the data was collected from the entire population, all coefficients are naturally statistically significant.

Relationship between number of pageviews (NPV) and quality classifications made by ORES (QC). The box-plot of NPV by QC (Fig. 2) shows that the distribution of NPV for each quality category is right-skewed, which hints at log-transforming NPV. This has an immediate advantage. Distributions of log-transformed NPV have similar variances and are symmetric (Fig. 5), which satisfies the assumption of and thus allows for modeling using a linear regression model. Fig. 5 also illustrates that median of log-transformed NPV increases consistently as quality classification improves. To emphasize this trend, Fig. 6 shows the increments of medians and means between neighboring quality categories. Both the increment of median and the increment of mean are comparatively small from “B” to “GA”.

Regression with QC as explanatory variable and log-transformed NPV as response variable reveals that QC is able to explain 29.4% variance in log-transformed NPV. Therefore, the QC of an article has a moderate predictive power for its log-transformed NPV. QC can explain 6.4% more variance in log-transformed NPV than log-transformed NEL and is therefore a slightly better predictor than log-transformed NEL. Regression results are shown in Table 1 and

Quality Classification	Estimate of Coefficient
Intercept (Stub)	5.94
Start	1.13
C	2.25
B	2.90
GA	2.96
FA	3.61

Table 1. Results of regression with quality classification as the explanatory variable and log-transformed number of page-views as the response variable. The R-squared value is 29.4%.

Quality Classification	Estimate of Coefficient
Intercept (Stub)	1.41
Start	0.61
C	1.60
B	2.28
GA	2.40
FA	2.92

Table 2. Results of regression with quality classification as the explanatory variable and log-transformed number of external links as the response variable. The R-squared value is 47.2 %.

all coefficients are statistically significant. Due to collinearity between QC and log-transformed NEL, a linear regression model that uses both QC and log-transformed NEL to predict log-transformed NPV has an R-squared value of 31.6%, which only exceeds 29.4% by a very small margin.

Discussion

The first research question inquired into the relationship between quality classification and number of external links. The hypothesis was that number of external links increases as quality classification improves, because both metrics are intended to measure the same underlying factor, that is, quality of articles. Analysis demonstrated that number of external links indeed increases as quality classification improves, except from “B” to “GA”. ORES’s wikipedia article states that ORES determines article quality by “looking at structural features and number of citations and do not access article writing quality, tone, standpoint and etc.” An R-squared value

of 47.2% shows that number of external links or citations plays an important role in determining quality classification made by ORES. Structural features of articles might explain the remaining variances.

Analysis has also resulted in some interesting observations. The fact that quality categories “B” and “GA” have almost identical median and mean number of external links is rather peculiar, given that there is one more quality category between “B” and “GA”. One conjecture is that articles in quality category “B” and articles in quality category “GA” differ in some other ways, but this requires further research for confirmation. One additional concern is how the original data was tagged and whether the tags were checked, because it can be very difficult for human administrators to classify articles into 7+ categories without a clear guideline of what each category represent. Future studies can look at article-quality evaluation metrics used by human administrators, data used to train ORES and the algorithms used by ORES to gain a deeper understanding of such peculiarity. A minor suggestion for future studies is that, instead of using number of external links, number of external links per 500 words may be a better indicator of article quality.

The second research question probed the relationship between quality classification and number of pageviews. Given the assumption that users are self-selective in what articles they read based on their quality, the number of pageviews should increase when quality classification improves. Analysis has confirmed this hypothesis. Despite the fact that this study is correlational, the good news is that quality classification made by ORES is a better predictor of the number of pageviews compared to the number of external links. This makes sense: structural clues are better visual indicators of article quality and thus help users better determine article quality.

A limitation of this analysis is that it is not informative of the underlying factors that contribute to this relationship. In other words, it is hard to quantify the degree to which quality classification made by ORES is a good reflection of article quality due to correlational nature of the study. A major confounding factor is that not all articles are created equal, i.e., some topics (e.g., commonly prescribed medicines) are searched more frequently than others (e.g., a rare genetic illness). It is quite possible that popular articles have more viewers initially and thus have more contributors and better quality. Future studies can control for such confounding variables and see whether the relationship persists and how does the strength of the new relationship compare to the strength of the relationship reported previously.

Based on the conclusion of the second research question, this study proposes a novel metric for prioritizing articles for moderation. Articles with low quality classifications but high numbers of pageviews are prioritized for moderation because they spread poor information to a large number of users. In other words, an article's priority is the distance between its number of pageviews and the mean number of pageviews for its quality category. Such a tool will also nicely fit into the workflow of human moderators on Wikipedia.

References

Hughes, B., Joshi, I., Lemonde, H., & Wareham, J. (2009). Junior physician's use of Web 2.0 for information seeking and medical education: a qualitative study. *International journal of medical informatics*, *78*(10), 645-655.

Yang, Zhihan. (2019). *Do more credible articles tend to have more views?*. Unpublished manuscript.