Looking for a Needle in a Haystack: Predicting Wikipedia Edits

Natalia Timakova, Nathaniel Weinman, Ugur Yildirim

(timakova, ugur.yildirim, nathaniel_weinman)@berkeley.edu

Abstract

Wikipedia is a widely-used public encyclopedia, getting over 200 million views per day [9]. However, it is regularly being updated with new information, more than 300 thousand times per day [9]. If it was possible to predict which articles were likely to be edited soon, Wikipedia could notify readers that there may soon be new information. Separately, models use Wikipedia edits as a feature to predict, for example, movie box office success [3]. Well-predicted near-future edits could be a relevant feature to models such as these, allowing them to identify trends slightly sooner.

We found that, on a balanced subset of edited and unedited articles, a Gradient Boosting model was the most effective compared to six other classifiers that we trained *(see Introduction)*. By using features from the main and talk namespaces for articles based on size, view count, edit count, and minor edit count, it was able to predict the probability of an article being edited with roughly 76% accuracy on a sampled dataset where half of articles were edited. Total views and current size of the article were the most significant features for the model, much more than trends in data.

Introduction

How many and what kind of features is enough to predict whether a Wikipedia article will be edited next? We started with a dataset containing over 20 features occurring "naturally" in Wikipedia, such as number of page views, both for namespace pages (i.e. articles) and talk pages which support a discussion about a particular article. In order to track a possible dynamic in views and edits, we split the 30-day period preceding a potential edit and retrieved numbers specifically for these time slots. We also generated new, "synthetic" features based on this initial set, effectively extending our feature space to over 90 features. Finally, we selected 50 those which gave us the best data representation, and trained seven classical ML classifiers on them:

1. Logistic Regression
2. Decision Tree
3. LASSO
4. k-Nearest Neighbors
5. Random Forest
6. Multi-Layer Perceptron
7. Gradient Boosting

We chose 50% accuracy on the balanced set (and 99.9% on imbalanced one) to be our baseline, which would be an output of a classifier consistently predicting only negative (i.e. unedited) class. All our models exceeded the baseline on the balanced data, but none reached the baseline on imbalanced data.

We found that the 3-5 most consistently important features were the total size of the article, the article's latest and average (for the entire 30-day period) sizes, and the article's total number of edits. Features describing the discussion dynamics on talk pages were shown to be less important.

We also set two hypotheses when we started our exploration:

1. The larger size of the article is correlated with a lower likelihood of it to be edited;
2. The bigger number of views of the article is correlated with a higher likelihood of it to be edited.

Our Logistic regression model supports both hypotheses (see Fig. 5).

Our error analysis suggests that our data is not linearly separable and that the difference in types of edits can be the core issue hindering the better classification score. Our literature research on Wikipedia editors' behavior indirectly supports this thesis.

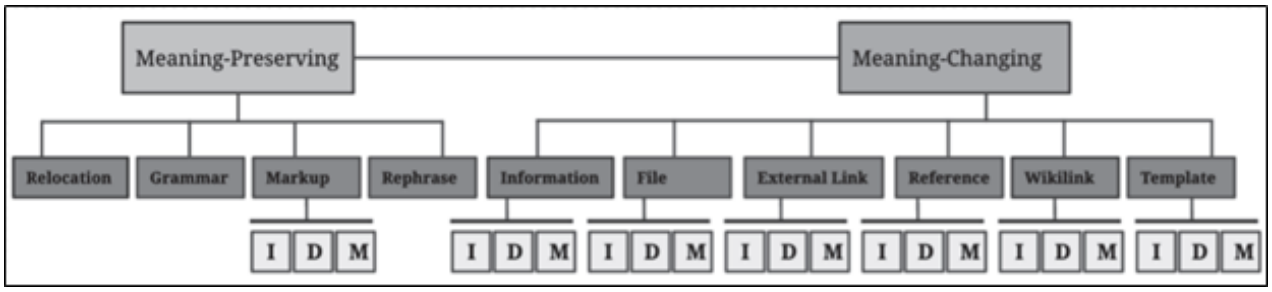

Figure 1: The Taxonomy of Edit Categories. Note: Insertion is abbreviated as I, Deletion as D and Modification as M. Diyi Yang et. al.[1]

Previous work

There are a few lines of research that use data on Wikipedia editors' activity and article popularity.

The first one focuses on editors' behavior and role identification. An example of such a research would be Diyi Yang et. al. [1], who proved that articles in different stages of quality require different types of editors, and there are eight types of such editors. What is especially relevant to our project, is the taxonomy of edit categories and a set of features Yang et. al. developed to classify an edit into one or more of the edit categories. Because "just as documents are mixtures of topics, editors are mixtures of roles," the authors use graphic model underlying the Latent Dirichlet Allocation (LDA) - method widely used for topic modeling. In the second section of the paper, the authors study how different editors' contributions influence article quality. They built 24 regression models to predict edit counts in each individual edit category from editors' role distribution.

The second group of research is focused on predicting certain events, like disease outbreaks [2] or movie box offices [3]. Twitter data is usually used along with Wikipedia in this line of research. It is extremely relevant for our project to account that real life events significantly impact editor's activity on Wikipedia, and even a completed article which haven't seen revisions for months can become trending among editors if a certain event triggers its popularity.

The third, and quite massive frontier of research is Wikipedia vandalism detection. It's relevant to our project because vandalism is a type of edit, and since certain articles are especially prone to attacks and attract a lot of damaging edits, we might want to take them into account in our analysis. For example, Kiesel et. al. [6] use edit timestamps to understand spatio-temporal characteristics of the revisions made by vandals. In particular, the authors discovered that vandalism is more likely on weekdays than on weekends, and is equally of characteristic for all seasons except summer. Their findings are based on visual inspection, backed by analysis of the variances of the average vandalism ratios (with Cohen's *d*) and the significance test (Welch Two Sample t-test).

Finally, the fourth group of whitepapers is focusing on article quality. For example, Blumenstock [7] used word count to successfully predict whether an article is among the featured ones (one of the signs of article quality). A simple binary classification technique - count every article over 2,000-words long as a featured one - helped reach 96.31% accuracy. A multi-layer perceptron yielded 97.15%, k-nearest neighbor - 96.94%, logit model - 96.74%, and a random-forest classifier - 95.80% accuracy.

Data

We leveraged two data sources for our analysis, both of which are built off information that Wikipedia provides publicly through APIs.

First, we used Wikimedia's Quarry interface [10], which provides an SQL interface into the MediaWiki database. The MediaWiki database provides metadata about Wikipedia articles (e.g. revision histories) from the beginning of 2016. Unfortunately, there are strict limitations on the output size of result tables through this interface. We were able to run queries by hand to get edit and minor edit counts per article and namespace, one day at a time.

To get more complete data, we used Page View Statistics for Wikimedia Projects [10]. This is a large, hourly data dump of view counts that is also provided by Wikimedia and also provides data from the beginning of 2016. We were able to write a Python script to download these hourly files, aggregate them into daily statistics, and write them to local CSVs. This datasource gave us view counts and size by article and namespace.

Combining these data sources allowed us to generate features based on article name, size, views, edits, and minor edits over a 30-day historic period from the Main and Talk namespaces. However, due to the limitations of Quarry, we were unable to include additional metadata for unedited articles.

Methodology

Initial Exploration

Preliminary analysis of our data shows that ~1/1000 articles viewed on a given day are edited. To handle the large imbalance of this data, we ran our analyses on a dataset of the edited articles and downsampled unedited articles. After removing entries with incomplete data, we were left with 82,461 observations to analyze for training (41,613 unedited, 40,848 edited).

Feature Generation

Previous work has shown that size, views, and edits are all interesting, descriptive features of Wikipedia. Though we didn't find related work to support this, we believed that time trends in these, as well as analogous characteristics of the talk pages for each article, may be relevant to our specific challenge of predicting edits. To represent time trends, we averaged data for the following four periods (without overlapping): from 30th to 8th day before prediction date, 7th to 4th, 3rd to 2nd, and for the day before the prediction date. To this end, we also generated several features from these metadata, representing change in average views and edits for the abovementioned periods. After some analysis of feature importances, we were left with:

- Current article size
- Average size for the entire 30-day period
- Change in average size in main and talk namespaces between the four consecutive periods (1 and 2-3rd day, 2-3rd and 4-7th, etc.)
- Total views in main and talk namespaces for the four periods
- Change in average daily views in main and talk namespaces between the four periods
- Total edits in main and talk namespaces on the four periods
- Change in average daily edits in main and talk namespaces between the four periods
- Total minor edits in main and talk namespaces for the four periods
- Change in average daily minor edits in main and talk namespaces between the four periods

Changes were represented both as an absolute difference and a modified ratio (adding 1 to numerator and denominator to avoid divide-by-zero errors). We did not explore making any simple transformation on the original data, such as analyzing log-scaled versions.

Feature Exploration

The full correlation matrix can be found in the appendix in our GitHub repo, but some interesting results are highlighted here. Surprisingly, the current size and average size over 30 days had only a .75 correlation, indicating that edits over

the 30 day period significantly affect article size. Total views over 30 days, total edits over 30 days, total minor edits over 30 days, and the ratio in minor edits between 4-7 and 2-3 days previous all have a 10% or greater correlation with our predicted variable, whether an article will be edited in the next 24 hours. The most recent size has only a ~6% correlation despite having the largest coefficient on a trained logistic regression model.

Models

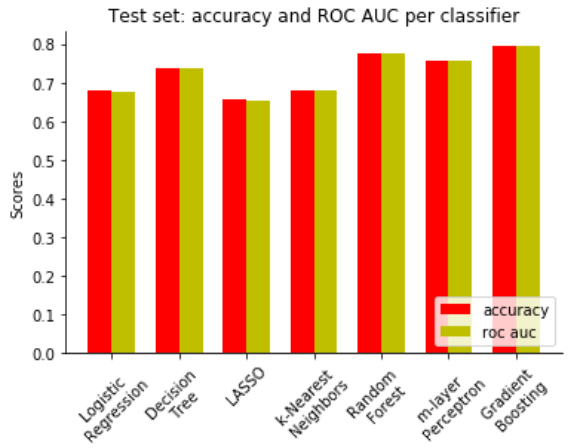

Figure 2: Average accuracy and ROC/AUC scores from cross-validation on training data

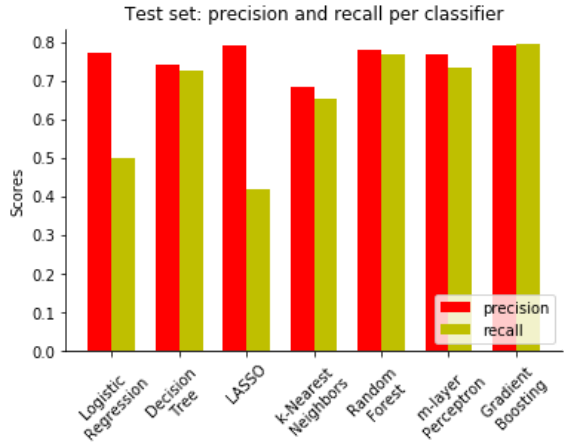

Figure 3: Precision and recall from cross-validation on training data

As we were focused on a binary classification problem, we explored ridge logistic regression, decision tree, LASSO logistic regression, k-nearest neighbors, random forest, multi-layer perceptron, and gradient boosting models. For each model, we used 5-way cross-validation on our test data to hand-tune hyperparameters based on average accuracy on predictions. Figure 2 shows the results of these models.

Both ridge and LASSO logistic regressions performed the worst. Their training accuracy is similarly low, indicating that our data is not easily linearly separable in log odds space with our features. As expected, decision trees performed slightly worse than random forest, which performed slightly worse than gradient boosting. Gradient boosting performed the best out of all our models, and better than the ensemble model (not shown here)[1].

Most of our models had higher precision than recall, as shown in Figure 3. Though we were focused primarily on accuracy, further research could inform the relative impact of precision vs. recall for our intended use cases.

Feature Importance

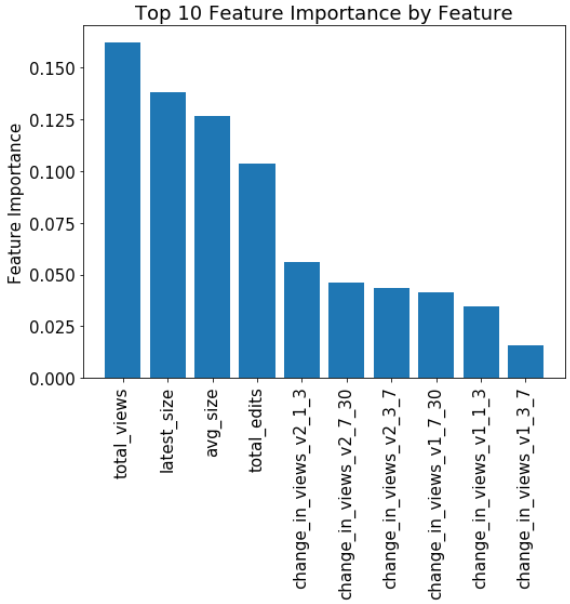

Figure 4: Top 10 features by importance for gradient boosting

[1] We built our ensemble using bagging method, which would work better if consisted of more "strong" than "weak" classification models.

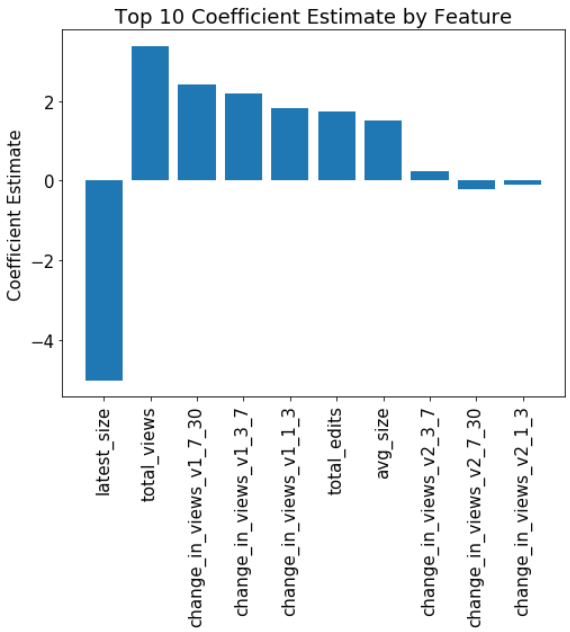

Figure 5: Top 10 features by importance for logistic regression

All of our models, similar to what is shown for gradient boosting in Figure 4, had total views over 30 days, current article size, average article size over 30 days, and total edits of 30 days in the 10 most important features, though the ordering differed between model.

In addition, Logistic regression coefficients (Fig. 5) demonstrated the directional correlation between edits and latests size as well as total views of an article. Both decreasing latest size and increasing total views correlate with a higher likelihood of an article to be edited.

Interestingly, none of the models found parameters from the talk namespace as particularly important. Also, against expectations, most models found current and total data over 30 days more important than recent trends in our features.

Error Analysis

Unlike other models, our gradient boosting model had nearly equal precision and recall.

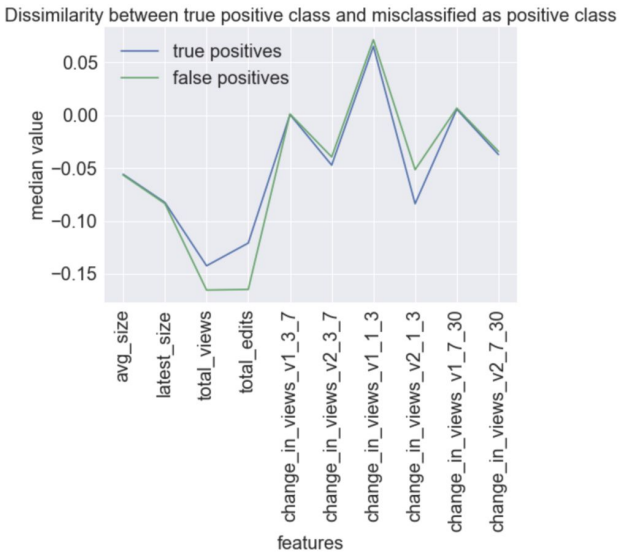

Figure 6: Median values of 10 most important features for true positives and false positives

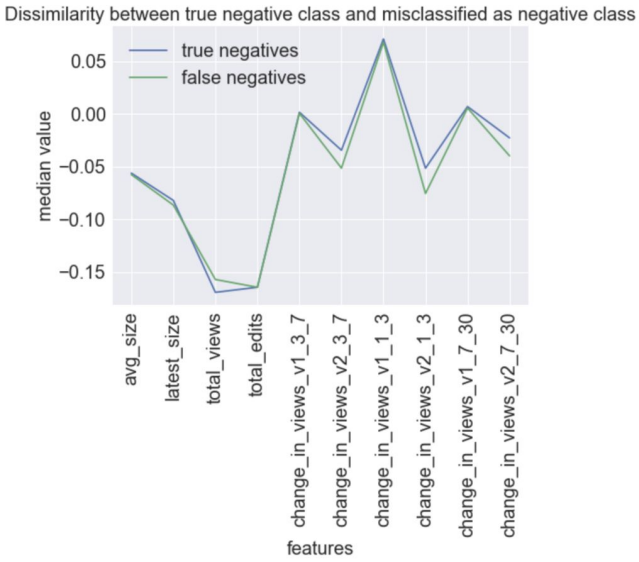

Figure 7: Median values of 10 most important features for true negatives and false negatives

Figures 6 and 7 visualize the median values of various features for our models. Though gradient boosting is not reliant on medians in the same way other models may be, we hoped this would indicate why the model was having trouble classifying certain articles. Figure 6 suggests that a better focus on total edits could be helpful at separating true and false positives. However, it

was already marked as the most important feature of our model, and, as shown by Figure 7, further weighting it would likely hurt the ability to separate true and false negatives. There is some hope for both misclassifications that a further weight on total views may have been able to help, though we have not yet explored this.

We did similar analysis based on minimum and maximum values of features, which can be seen in the appendix in our GitHub repo, but that also did not give us obvious ways to improve our predictions.

Results

Model	Gradient Boosting	Logistic Regression
Accuracy	.755	.732
ROC/AUC	.753	.730
Precision	.923	.979
Recall	.552	.470

Table 1: Results from Gradient Boosting and Logistic Regression models on down-sampled balanced test data

Model	Gradient Boosting	Logistic Regression
Accuracy	.028	.020
ROC/AUC	.427	.476
Precision	.0007	.0008
Recall	.827	.932

Table 2: Results from Gradient Boosting and Logistic Regression models on imbalanced test data

Tables 1 and 2 shows the results of our final gradient boosting model (our expected best model) as well as our logistic regression model (our expected worst model). Both models were trained on balanced data. The gradient boosting model performed reasonably well on the balanced data set, but expectedly fell very short of the 99.9% baseline of always predicting no edits on the imbalanced data set.

Interestingly, when trained on balanced data, the ridge logistic regression model did better on predicting imbalanced data than gradient boosting in terms of ROC/AUC, precision, and recall despite being the weakest model on the balanced data. This indicates it was stronger at correctly classifying true negatives, which in this case is a strong majority class. This may be because decision trees (and therefore gradient boosting models) are more able to overfit on the balanced data on the positive class, meaning they'd do worse at finding larger numbers of negative classifications.

Future work

Our analyses were primarily trained and optimized for the balanced data set. However, for these predictions to be useful, they must perform reasonably well on imbalanced data in-the-wild. We believe we could significantly improve our model by leveraging Machine Learning techniques targeting imbalanced data, for example informed undersampling [8].

We would also like to include features from more sources, as related work indicates these could be meaningful predictors. For example, we'd like to include topology of inter-article linking [5], and features based on applying natural language processing techniques to article and edit content [2].

We were also discussing the possibility to include news trends from platforms such as Twitter [5] into our models. Although this idea looks reasonable, we realized the underlying assumption - that Wikipedia community and Twitter community are essentially the same people or at least very similar in their interests - could not

hold. Which means the signal we would get from trending event on Twitter could be misleading.

Another feature worth exploring is the day of the week and the time of the day the edit is anticipated on. At least one research paper shows some types of editorial activity can be linked to these characteristics [6].

We would also like to collaborate with Wikimedia Foundation to create easier access to the rich data they have. Other researchers, such as Kämpf et al. [4], have also found difficulty accessing relevant data. Though they attempted to create an open source platform to improve access, we were unable to find one that was not password protected. Easier access to Wikipedia metadata could be very valuable to other researchers in the future.

Conclusion

Successfully predicting near-future edits to a Wikipedia page would be valuable both as a user experience and as a feature into other models. Based on size, view, and edit features, our model was able to do a reasonably good job predicting edit likelihood on a down-sampled, balanced dataset. More work must be done for it to be reliable in the wild. Fortunately, there are sets of techniques focused on imbalanced data that could help close this gap.

Size and views were generally the most important features in this model. Interestingly, data on the talk namespace was not very valuable to our models. Additionally, directional trends (either in absolute difference or relative ratio) were not very valuable, indicating that much of the necessary data is in the current state and very recent past. Some other features may also be helpful in improving our edit frequency prediction.

Lastly, it would be even more valuable if a model was able to predict the type of edit that would occur, for example following Yang et. al's classification [1]. We realized that different types of edits may have different patterns reflected in features. For example, when the Wikipedia community works on a new article, the dynamic in edits may be more telling, than number of views; while when the anticipated edit is an act of vandalism, the preceding number of views should be more important.

Our work provides a first step in trying to better predict and understand Wikipedia edits.

Our models and data can be found at https://github.com/uguryi/info-251-fall-2017-final-project or through the UC Berkeley School of Information project gallery at https://www.ischool.berkeley.edu/projects/2017/looking-needle-haystack-predicting-wikipedia-edits.

References

1. Diyi Yang, Aaron Halfaker, Robert Kraut, Eduard Hovy. Who Did What: Editor Role Identification in Wikipedia. Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)

2. Reid Priedhorsky, Dave Osthus, Ashlynn R. Daughton, Kelly R. Moran, Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle. Measuring Global Disease with Wikipedia: Success, Failure, and a Research Agenda. CSCW '17 Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, Portland, Oregon, USA — February 25 - March 01, 2017

3. Márton Mestyán, Taha Yasseri, János Kertész. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. PLOS ONE, Volume 8, Issue 8, August 2013

4. Mirko Kämpf, Eric Tessenow, Dror Y. Kenett, Jan W. Kantelhardt. The Detection of Emerging Trends Using Wikipedia Traffic Data and Context Networks. PLOS ONE, December 31, 2015

5. Stewart Whiting, Joemon M. Jose, Omar Alonso. Wikipedia as a Time Machine. WWW '14 Companion Proceedings of the 23rd International Conference on World Wide Web, Pages 857-862, Seoul, Korea — April 07 - 11, 2014

6. Johannes Kiesel, Martin Potthast, Matthias Hagen and Benno Stein. Spatio-temporal Analysis of Reverted Wikipedia Edits. Association for the Advancement of Artificial Intelligence, 2017

7. Joshua E. Blumenstock. Size Matters: Word Count as a Measure of Quality on Wikipedia. WWW 2008 / Poster Paper April 21-25, 2008 · Beijing, China
8. Haibo He, Edwardo A. Garcia. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering (Volume: 21, Issue: 9, Sept. 2009)
9. Statistics:
- https://analytics.wikimedia.org/dashboards/vital-signs/#projects=enwiki/metrics=Pageviews
- https://stats.wikimedia.org/EN/TablesDatabaseEdits.htm

10. Data sources:
- https://quarry.wmflabs.org/
- https://dumps.wikimedia.org/other/pagecounts-raw/