Learning from history: Predicting reverted work at the word level in Wikipedia

Jeffrey M. Rzeszotarski, Aniket Kittur
Human-Computer Interaction Institute
Carnegie Mellon Univerity
Pittsburgh, PA
{jeffrz, nkittur}@cs.cmu.edu

ABSTRACT

Wikipedia's remarkable success in aggregating millions of contributions can pose a challenge for current editors, whose hard work may be reverted unless they understand and follow established norms, policies, and decisions and avoid contentious or proscribed terms. We present a machine learning model for predicting whether a contribution will be reverted based on word level features. Unlike previous models relying on editor-level characteristics, our model can make accurate predictions based only on the words a contribution changes. A key advantage of the model is that it can provide feedback on not only whether a contribution is likely to be rejected, but also the particular words that are likely to be controversial, enabling new forms of intelligent interfaces and visualizations. We examine the performance of the model across a variety of Wikipedia articles.

Author Keywords

Wikipedia, Applied Machine Learning, Reverts

ACM Classification Keywords

H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces.

INTRODUCTION

Wikipedia is one of the largest and most successful examples of online social production, gathering hundreds of millions of contributions into an encyclopedia of human knowledge that is among the top ten most accessed websites in the world. However, its very success in can pose an intimidating challenge to editors who may need to sift through hundreds or thousands of contributions to a page to avoid rekindling previous conflict, rehashing previous decisions, or violating agreed upon norms or policies. If they fail to do so, their work may be discarded or "reverted" with other editors undoing their work. This type of action can be discouraging and punitive, since an editor's contributions -- often made in good faith -- are discarded wholesale by the community. This may be especially true for newcomers, who have been reverted at increasingly high rates over the past years [11]; even for veteran users the experience of being reverted can be damaging to future participation [4,5].

Since every contribution is saved, in theory an editor could simply inspect the edit history for reverted contributions, vandalism, conflict, article conventions, and administrator caprices, all of which could be useful in avoiding future reverts. However, for many pages the article history is prohibitively large. In the case of contentious articles like *Abortion* and *Scientology*, there are over ten thousand edits, totaling hundreds of megabytes of text. Compared to the plain text of Jane Austen's *Pride and Prejudice*, the complete history of the *Abortion* article is roughly one thousand times larger, and a summary of word-by-word changes twenty times [2]. Few editors are likely to expend the effort of thoroughly learning a page's history in order to produce better edits.

However, this same history, processed appropriately, has the potential to help editors understand which contributions would be appropriate and valued. This paper describes a method of identifying edits likely to be rejected using machine learning on word level features in the edit history of an article. By learning from the past history of an article, the model not only identifies edits that are likely to be reverted, but also provides feedback at the word level. This word-specific information can enable new forms of page-specific visualization and feedback for newcomers and experienced editors alike.

STUDYING REVERTS

Reverts are employed by Wikipedians for a variety of reasons. Many reverts are used to quash vandalism, a constant problem in an open online encyclopedia. Researchers have developed means of identifying potential malicious edits, including examining macro scale features of edits, evaluating text case, vulgarity dictionaries, and user characteristics such as edit counts [1,8]. Vandals are often anonymous, and wipe entire pages or replace large swaths of text, as reflected by the features in vandal identification models [8]. These models do not examine the entire edit contents word-by-word, instead utilizing more general features about users and edit characteristics. Other models do include words, but merge edits together [3]. In both cases, these methods can overlook valuable per-article vandalism behaviors, such as words that are vandalistic in one article context (the word "gay" used as a pejorative), but not in another (the word "gay" in an article about LGBT rights), that an article-level word model could discover.

Furthermore, we are interested in going beyond predicting simple vandalism to understanding differences in the perceived value of contributions. For example, reverts are also employed by Wikipedians as a part of inter-editor conflict [7]. Editors claim strong ownership over edits or articles, or guard pages based on their orthodoxy, resulting in 'revert wars' where cultural, historical, or other biases are repeatedly contributed and undone [5,7]. Much work has gone into visualizing and investigating this conflict using editor level features (such as the persistence of their edits or who they revert); however, these techniques do not provide editors with concrete feedback on improving their contributions [11]. While conflict is inevitable in a social space as rich as Wikipedia, newcomers may stumble into a warzone and never return. Word-by-word models could identify potential 'battlezone' words and help newcomers avoid them until they better understand the article's history.

Furthermore, many reverts are directed towards good faith edits, often made by newcomers, that are not perceived as valuable. This may be because they violate one of the many precepts contained in dozens of pages of Wikipedia policy, or per-article rules and conventions. For instance, the article on *Abortion* has a multiple-page discussion about editor conventions for the use of the terms "termination", "death", and "murder" in editing. Since such rules are held in pages-long discussion threads, the cost of understanding this level of nuance is high. Much like in a conflicted article, a word-by-word model could identify these word level conventions and warn newcomers or direct them to places to learn more.

MODELING ARTICLE HISTORIES

Our model operates on a per-article basis to gather features that are contextual to each Wikipedia page. This facilitates nuanced judgments based not only on general Wikipedia policy, but also article-specific conventions and conflicts. We use the contents of edits rather than user features since just as past vandals might make a good faith edit, a newcomer might make mistakes, and can be applied to contributions even by editors with no editing history.

To build a word-by-word model of reverting behavior for a given article on Wikipedia, we extract the edit history of the article using the Wikipedia API. These edit histories include full copies of each article for every edit made, editor comments, and usernames. Based on the number of edits, such histories range in size from several megabytes up to a gigabyte of plain text because complete copies are stored for every contribution. As a result, this process can be time consuming because the live Wikipedia API is not well adapted for dumping entire article histories.

Once we have a complete article history composed of many copies of the article evolving as editors make changes, we tokenize each contribution according to Wikipedia syntax into word level tokens. Since the Wikipedia servers store edits in code markup, our tokenizer checks for special cases such as references, links, tables, and images and counts such elements as 'words' as well. Once we have tokenized the complete article history, we conduct a text comparison (diff) of consecutive tokenized edits by temporarily converting each token into a unique Unicode character and performing a character level text comparison. This captures word level changes rather than character level ones. For example, a comparison of the phrase 'brown fox' -> 'brown foxes' is recorded as removing the word 'fox' and adding 'foxes' rather than inserting the letters 'es'. This procedure captures the region in which an editor is making changes over the long history of an article, providing a middle ground between sentence level diffs that may be too high level or letter-by-letter diffs that would focus on minutiae.

Once we have determined through text comparison how each editor changed the article in every edit, we convert their changes into datapoints. We count the number of times each word is added or removed as two separate features, leading to feature spaces that are on average three to ten thousand words in size. We also include comment length, the anonymity of the editor, and his or her edit count and time registered on Wikipedia. We identify reverted work using the same method as previous researchers [7,9,10] of MD5 hashing the contents of the page at each edit and looking for pages that hash the same, indicating a new edit has the same contents as a previous iteration of the page. We place the datapoints into two classes based on whether hashing shows that the changes have later been reverted.

CLASSIFYING WIKIPEDIA EDITS

To test the accuracy of our model and the quality of its feedback, we randomly sampled 150 articles from Wikipedia that each had over 1,000 contributions (well over 10,000 articles have at least this many edits). We chose 1,000 edits as a benchmark for a well established Wikipedia page that is not only likely to receive more future edits, but also is likely to have a varied and rich history. Our sample varied from as few as 1,114 edits to as high as 10,593 edits

Classifier	*Accuracy*	*F-Measure*	*Area Under ROC Curve*
Naïve Bayes	81.2% (SD 9.7%)	0.767 (SD 0.100)	0.659 (SD 0.118)
SMO Support Vector Machine	86.9% (SD 3.0%)	0.873 (SD 0.029)	0.802 (SD 0.056)
Random Decision Tree Forest	89.9% (SD 2.6%)	0.892 (SD 0.026)	0.880 (SD 0.052)

Table 1: Mean classifier accuracy, F-measure, and area under ROC curve evaluated using 150 randomly sampled Wikipedia articles under 10-fold crossvalidation

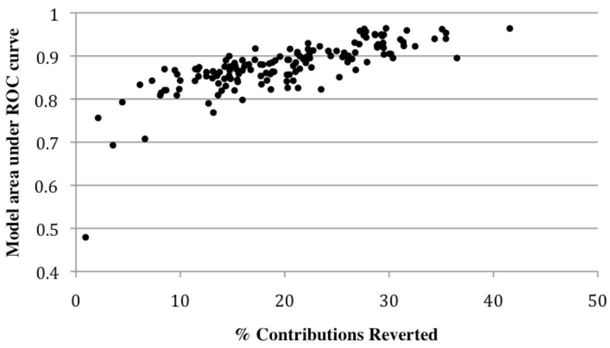

Figure 1: Proportion of edits that were reverted (x) in a given article versus the area under the ROC curve for its model (y).

(Mean 3,149 , SD 1,716), with as little as 11 reverted contributions and as many as 3,241 reverts (Mean 657, SD 496). Reverts comprised as little as 0.9% of article histories and as much as 41.6% (Mean 20.0%, SD 7.9%).

We used the Weka toolkit to construct and evaluate models on a per-article basis since each article represents a very different spectrum of editor behavior. We built and tested models using naïve Bayes classifiers (proven to work well on 'bag of words' models), support vector machines using sequential minimization optimization (able to deal with a large featureset), and decision tree forests (well adapted to highly unbalanced classes). We also used costs to even out the sometimes strongly unbalanced classes for the support vector model, and feature selection to limit the large number of extraneous words/features. We further limited the feature set by removing words/features that were shorter than 4 characters long. Overall, while naïve bayes models performed fairly, support vector machines and decision tree forests performed comparably well in a random test sample of 50 articles. Random decision tree forests proved to generate the best classifiers (Table 1), and were used to evaluate all 150 articles.

Using random decision tree forests under 10-fold cross validation, word level models of edits predicted which contributions were likely to be reverted with generally high accuracy, with a mean accuracy of 89.4% (SD 2.5%), an average F-measure of 0.886 (SD 0.026), and an average area under the ROC of 0.876 (SD 0.058). Several articles did not fare as well, but those primarily did not have a significant number of reverts. For instance, one article model, *RahXephon*, achieved a ROC area of only 0.69, but reverts only comprised 3.5% of its 1,693 edits. In general, fewer reverts corresponded significantly with areas under the ROC closer to 0.5 (an indicator of poor performance) (F(149)=199.7, p<0.001, Figure 1). We also investigated models on content with less than 100 edits such as the article on *Alkahest*, but were not able to predict reverts accurately, suggesting that this technique is limited to pages with a sizable number contributions. However, since such pages are more amenable to manual inspection, the need for a machine learning model is correspondingly lower.

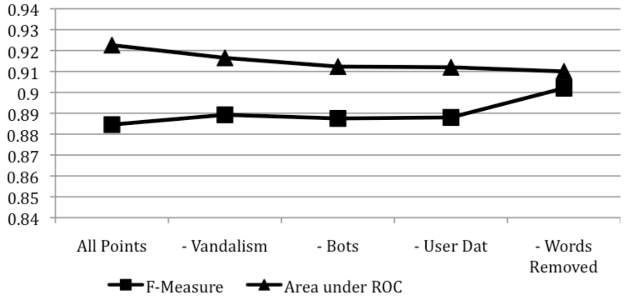

Figure 2: *Genetic Engineering* classification results as points/features are progressively removed

To explore the utility of different features we filtered out various points in the data. We filtered vandalism by checking for reverts with 'rvv' or 'vandal' included, and we filtered bots using the Wikipedia convention of the word 'bot' at the beginning or end of a user name. We also investigated eliminating features, including data about the user who made the edit, as well as limiting features to only words added (vs. removed). Using 10-fold cross-validation at each step, we investigated progressively eliminating features in the model using the *Genetic Engineering* article (see Figure 2). In the base case, the model showed high accuracy when including added and removed words, as well as vandalism, bots, and user data. Importantly, the model still obtained high accuracy when vandalistic edits and bots were filtered out, demonstrating its ability to identify meaningful value rather than just simple vandalism. Eliminating features for the words that a reverted edit removed, and only looking at words the edit contributed, the model still achieved high accuracy. We removed user data as well, limiting the model to knowledge of the user's anonymity, their comment length, and their contribution, with similar levels of accuracy. Finally, we cut the number of features in half by only examining words an editor added, still obtaining reasonable results. This suggests that word-level models may be robust across a variety of feature spaces.

IDENTIFYING WORD LEVEL CHARACTERISTICS

After verifying that the model accurately identified reverted edits, we extracted the individual normalized feature weights from a support vector classification model built out of the entire article history in order to examine word-by-word what edit characteristics lead to valued or discarded work. We coded a sample of various weight levels in the model.

We extracted the word feature weights from the *Genetic Engineering* SVM model, took a stratified sample of the words, and coded them based on Wikipedia policy and conventions (see Table 2). The features that were weighted highly towards reverted work (positive numbers) were generally those that violated policy or article conventions, had spelling errors, or had Wiki syntax errors. Neutral model weights could be construed as violating policy, but could also be employed otherwise. For example,

Word	Weight	Explanation
hello	2.00	"Hello" does not relate to the contents of the article
significant	0.09	"Significant" can be used with citations, such as, "The treatment showed a significant increase in mobility," or as a Wikipedia weasel word in "Significant numbers of researchers say genetic engineering is evil"
virus	-0.52	"Virus" is primarily neutral, used in sentences such as "The virus can be used as a vector…"

Table 1: Example Model Words, Weights, and Explanations for the *Genetic Engineering* article

"predicted" can be used in a useful phrase like "Mendel predicted", or in a phrase that violates Wikipedia 'weasel word' policies like "many scientists predicted". Low weighted (rarely reverted) words were often domain-specific to genetics. There were also many words that had no clear judgment for or against reverting throughout the weights, suggesting that there is a degree of noise in the data.

IMPLICATIONS FOR DESIGN

There are many potential design implications for the use of word level modeling of reverts. Foremost, we can immediately apply this classifier in the editing process to provide feedback to newcomers as they edit. The model could inform editors when their edit is likely to be reverted, enabling them to reflect on and revise their contribution to increase its perceived value. Coupled with heuristics that identify common reasons for being reverted, such as Wikipedia policy and discussion page mining, the feedback could be even more effective in improving both contribution quality and retention, as editors whose work is reverted are less likely to stick around [6].

The word-by-word model weights might also be used visually to show problem areas in the article, heatmapping sections that are suspect because they are more contentious. Word ratings between articles could be compared, showing the difference in editing conventions between articles. One could also calculate a set of weights for all Wikipedia pages combined, gathering information about general trends.

It is also possible that word level models might be applicable in other online settings, including online forums and reviews. There, content can be flagged as either offensive or of low value. There is potential for similar classifier models designed to help contributors write higher quality content.

CONCLUSION

Our research suggests that there is indeed a great amount of information encoded by discarded work on Wikipedia at the word level that could be useful for improving sensemaking and contribution. Using only the word-by-word changes made by editors discarding work, we were able to predict future discarded work, and characterize the nature of what types of words are reverted in an article, even when filtering out simple vandalism and bot edits. This suggests that there is great potential for future work using word level revert features to provide feedback for newcomers and experienced users about their edits.

ACKNOWLEDGMENTS

We thank the Carnegie Mellon Social Computing Lab for their support. This research was supported by NSF grants OCI-09-43148, IIS-0968484, IIS-1111124 and a grant from Carnegie Mellon's Center for the Future of Work.

REFERENCES

1. Adler, B., Alfaro, L. de, and Pye, I. Detecting Wikipedia Vandalism using WikiTrust. *Notebook Papers of CLEF*, 2010, 22–23.
2. Austen, J. Pride and Prejudice. *Project Gutenberg*, 1813
3. Druck, G., Miklau, G., and McCallum, A. Learning to predict the quality of contributions to wikipedia. *In Proc. AAAI Workshop on Wikipedia and Artificial Intelligence*, 2008, 7–12.
4. Ekstrand, M. and Riedl, J. rv you're dumb: identifying discarded work in Wiki article history. *In Proc. Symposium on Wikis and Open Collaboration*, 2009, 4.
5. Halfaker, A., Kittur, A., Kraut, R., and Riedl, J. A jury of your peers: quality, experience and ownership in Wikipedia. *In Proc. Symposium on Wikis and Open Collaboration*, 2009, 1–10.
6. Halfaker, A., Kittur, A., Riedl, J. Don't bite the newbies: Revertings effect on the quantity and quality of Wikipedia work. *In Proc. Symposium on Wikis and Open Collaboration*. 2011
7. Kittur, A., Suh, B., Pendleton, B.A., and Chi, E.H. He says, she says: Conflict and coordination in Wikipedia. *In Proc. SIGCHI conference on Human factors in computing systems*, 2007, 453–462.
8. Potthast, M., Stein, B., and Gerling, R. Automatic vandalism detection in Wikipedia. *Advances in Information Retrieval*, 2008, 663–668.
9. Priedhorsky, R., Chen, J., Lam, S.T.K., Panciera, K., Terveen, L., and Riedl, J. Creating, destroying, and restoring value in Wikipedia. *In Proc. CSCW*, 2007, 259–268.
10. Suh, B., Chi, E., Pendleton, B.A., and Kittur, A. Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations. *In Proc. Visual Analytics Science and Technology*, 2007, 163–170.
11. Suh, B., Convertino, G., Chi, E., and Pirolli, P. The singularity is not near: slowing growth of Wikipedia. *In Proc. Symposium on Wikis and Open Collaboration* 2009, 1–10