DePP: A System for Detecting Pages to Protect in Wikipedia

Kelsey Suyehira
Computer Science Department
Boise State University
Boise, ID 83725 USA
kelseysuyehira@u.boisestate.edu

Francesca Spezzano
Computer Science Department
Boise State University
Boise, ID 83725 USA
francescaspezzano@boisestate.edu

ABSTRACT

Wikipedia is based on the idea that anyone can make edits to the website in order to create reliable and crowd-sourced content. Yet with the cover of internet anonymity, some users make changes to the website that do not align with Wikipedia's intended uses. For this reason, Wikipedia allows for some pages of the website to become protected, where only certain users can make revisions to the page. This allows administrators to protect pages from vandalism, libel, and edit wars. However, with over five million pages on Wikipedia, it is impossible for administrators to monitor all pages and *manually* enforce page protection. In this paper we consider for the *first* time the problem of deciding whether a page should be protected or not in a collaborative environment such as Wikipedia. We formulate the problem as a binary classification task and propose a novel set of features to decide which pages to protect based on (i) users page revision behavior and (ii) page categories. We tested our system, called `DePP`, on a new dataset we built consisting of 13.6K pages (half protected and half unprotected) and 1.9M edits. Experimental results show that `DePP` reaches 93.24% classification accuracy and significantly improves over baselines.

1. INTRODUCTION

In order to create reliable and crowd-sourced content, Wikipedia is based on the idea that anyone can make edits to the website. Although in certain circumstances, this is not the case. Some restrictions may be placed on pages because of an identified likelihood of damage if the page is left open for editing by anyone. Placing these restrictions is known as *page protection*[1]. There are different levels of page protection for which different levels of users can make edits (or, in general, perform actions on the page): fully protected pages can be edited (or moved) only by administrators, semi-protected pages can be edited only by autoconfirmed users, while move protection does not allow pages to be moved to a new title, except by an administrator. Page protections can also be set for different amounts of time, including 24 or 36 hours, or indefinitely. Examples of protected pages on English Wikipedia are *Drug* and *Bigfoot*. Users can recognize these kind of pages by the image of a lock in the upper right hand corner of the page.

Common motivations that an administrative user may have in protecting a page include consistent vandalism or libel from one or more users. A page protection can also be applied if edit wars are occurring. According to Wikipedia, an edit war is when two users cannot agree on the content of an article and one user repeatedly reverts the other's edits[2]. If a page becomes protected, this forces the users to go to the article's talk page in order to settle the conflict or to find help in resolving the dispute.

Currently, English Wikipedia contains over five million pages. Only a small percentage of those pages are currently protected, about 0.2 percent. Since many of the protected pages can be templates or user pages, the amount of protected article pages is even smaller. However, around 17 pages become protected every day[3]. This ratio makes it difficult for administrative users to monitor over all Wikipedia pages to determine if any need to be protected. Users can request pages to be protected or unprotected but an administrative user would have to analyze the page to determine if it should be protected, what level of protection to give, and for how long the protection should last, if not indefinitely. All this work is currently *manually* done by administrators.

In this paper we study for the *first* time the problem of deciding whether a page should be protected or not in a collaborative environment such as Wikipedia. Our contributions are the following. (1) We formulate the *new* problem as a binary classification task and propose a novel set of features to identify pages to protect based on (i) users behavior in editing page revisions and (ii) page categories. (2) We built a new dataset consisting of 13.6K pages (half protected and half unprotected) and 1.9M edits to test our approach. (3) We tested our features with different classification algorithms and we show that we are able to classify if a given article should be protected with an accuracy of 93.24%, significantly outperforming over baselines.

2. RELATED WORK

To the best of our knowledge, little research has been done

[1] https://en.wikipedia.org/wiki/Wikipedia:Protection_policy

[2] https://en.wikipedia.org/wiki/Wikipedia:Edit_warring

[3] We have observed the number of protected pages that are listed on the Wikipedia website from May 6, 2016 through Aug 6, 2016.

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983914

on the topic of page protection in Wikipedia. Page protection is a form of organizational "wiki-work". A few papers have researched different forms of wiki-work, including the featured article process [10], or mining editors' activity to automatically assign Barnstars [7]. Hill and Shaw [4] studied the impact of page protection on user patterns of editing. They also created a dataset (they admit it may not be complete) of protected pages to perform their analysis.

One of the main reasons to apply page protection is to prevent edit warring. This often happens on pages about controversial topics. Roitman et al. [9] proposed an algorithm to retrieve relevant articles containing claims for a given controversial topic. However, their algorithm requires the list of controversial topics as input, and no study has been done between retrieved pages and page protection.

Beside edit warring, another reason why pages are protected is to avoid or block vandalism from occurring. The problem of vandalism detection on Wikipedia has been heavily studied. Many works have attempted to predict which pages have been vandalized (see [3] for a survey) or when certain acts of damage have occurred [8, 6], while other recent research proposes an early warning system to detect vandal users [5]. Currently, ClueBot NG [1] and STiki [2] are the state-of-the-art tools used by Wikipedia to detect vandalism. ClueBot NG is a bot based on an artificial neural network which scores edits and reverts the worst-scoring edits. STiki is an intelligent routing tool which suggests potential vandalism to humans for definitive classification. It works by scoring edits on the basis of metadata and reverts, and computing a reputation score for each user.

There are not currently bots on Wikipedia that can search for pages that may need to be protected. WikiMedia does have a script[4] available in which administrative users can protect a set of pages all at once. However, this program requires that the user supply the pages, or the category of pages to be protected and is only intended for protecting a large group of pages at once. There are some bots on Wikipedia that can help with some of the wiki-work that goes along with protecting or removing page protection. This includes adding or removing a template to a page that is marked as protected or no longer marked as protected. These bots can automatically update templates if a page protection has expired. We believe the work done in this paper would be helpful to detect pages to protect by reducing the time many administrators have to spend checking and marking pages as protected and updating the page template, or the time other users have to spend watching pages instead of creating quality content.

3. DATASET

Wikipedia has different types of pages that may need to be protected, besides article pages. These include user pages, talk pages, and templates. Moreover, it has different mechanisms to protect a page as edit protection, move protection, or upload protection. As also stated in [4], building a dataset of protected pages is challenging because of inconsistencies in the Wikipedia data in reporting protection events. Thus, in this paper we concentrated on articles pages because these are more likely to have changes to the protection, and we consider the (semi and full) edit protection level as edit protection times can be easily retrieved from Wikipedia API.

As the percentage of protected pages, out of the total number of articles on Wikipedia is very small, we built a balanced dataset of protected/unprotected pages as follows.

We first collected the entire list of edit protected articles up to April 7, 2016 resulting in 6,799 pages. Then, we gather a list of unprotected pages of almost the same size by requesting random article pages from the Wikipedia API (we double checked if the random page was protected, and if so, we discarded the page). We collected a total of 6,824 unprotected pages.

For each protected/unprotected selected page, we then gathered up to the last 500 most recent revisions, if there were that many. However, if the page was protected, we only gathered the revisions up until the most recent protection. If there was more than one recent protection, we gathered the revision information between the two protections. This allowed us to focus on the revisions leading up to the most recent page protection. Revision information that we collected included the user who made the revision, the timestamp of the revision, the size of the revision, the categories of the page, and any comments, tags or flags associated with the revision.

Our final dataset[5] consists of a total of 13.6K (half protected and half unprotected) pages and 1.9M edits.

4. THE DEPP SYSTEM

In this section we describe the set of features we define to develop our `DePP` system. These features take into account the page revision history and the page categories.

The first group of six features is computed, for each page, on the whole edit history we have in our dataset (up to 500 edits per page) and looks at how Wikipedia users contribute to a particular page:

[E1] *Total average time between revisions*: pages that have very few edits over a long period of time are less likely to become protected (as their content is more stable) than pages with many edits that happen with little time between them.

[E2] *Total number of users making 5 or more revisions*: this feature counts the number of users who make more than five edits to a page.

[E3] *Total average number of revisions per user*: if there are many users making a few changes to a page, it is less likely to become protected than if a few users are making a lot of changes to a page.

[E4] *Total number of revisions by non-registered users*: this feature measures the number of changes made to a page from non-registered users. If a user has not spent the time to set up an account, it is less likely that they are a proficient user and more likely to be a spammer or vandal. Therefore, the more non-registered users that are editing a page, the more likely it is that the page may need to be protected.

[E5] *Total number of revisions made from mobile device*: similar to feature E4, this feature looks at the number of revisions that are tagged as coming from a mobile device. This is a useful feature because users making changes from a mobile device are not likely to be sitting down to spend time making revisions to a page that would add a lot of value. It is possible that a user making a change from a mobile device is only adding non-useful information, vandalizing a page, or reverting vandalism that needs to be removed immediately.

[4] https://www.mediawiki.org/wiki/Manual:Pywikibot/protect.py

[5] The dataset is available at https://sites.google.com/site/francescaspezzano/publications/depp

[E6] *Total average size of revisions*: it is possible that users vandalizing a page, or adding non-useful information would make an edit that is smaller in size. This is opposed to a proficient user who may be adding a large amount of new content to a page. For this reason, we measure the average size of an edit. Small edits to a page may lead to a page becoming protected more than large edits would.

In the second group of features we take into account the page editing pattern over the time. We define these features by leveraging the features E1-E6 as follows. For each page we consider the edits made in the latest 10 weeks and we split this time interval into time frames of two weeks (last two weeks, second last two weeks, etc.). Then, we compute features E1 to E6 within each time frame. This produces a total of 30 new features: 6 features times 5 time intervals. These features are denoted as Ei-j^{th} where $1 \leq i \leq 6$ identifies the feature and $1 \leq j \leq 5$ refers to the time interval. For instance E2-4^{th} represents the feature E2 computed in the 4^{th} last two weeks. The idea of these features is to monitor features E1-E6 over time to see if some anomaly starts to happen at some point. For instance, if a page is new we may observe a lot of edits of larger size in a short time after the page is created as users are building the content of the page. Later when the content is stable, we may observe fewer edits of smaller size representing small changes in the page. On the other hand, if the content of the page was stable and suddenly we observe a lot of edits from many users, it may indicate the page topic became controversial and the page may need protection.

The next features we propose use information about page categories[6]:

[NC] *Number of categories*: this feature counts the number of categories that a page is marked under. Pages that are classified under many different categories are likely to be more complex than pages that are grouped into fewer categories. They are therefore more likely to become protected.

[PC] *Probability of protecting the page given its categories*: given all the pages in the training set T and a page category c, we compute the probability $\texttt{pr}(c)$ that pages in category c are protected as the number of protected pages in T divided by the total number of pages in T having category c. Then, given a page p having categories $c_1, \ldots, c_n$, we compute this feature as the probability that the page is in at least one category whose pages have a high probability to be protected as

$$PC(p) = 1 - \prod_{i=1}^{n}(1 - \texttt{pr}(c_i))$$

In addition to the above two features, we define another group of features that shows how much features E1-E6 vary for a page p w.r.t. the average of these values among all the pages in the same categories as p. Specifically, given the set of pages in the training set T, we computed the set C of the top-100 most frequent categories. Additionally, for each category $c \in C$, we averaged the features E1-E6 among all the pages (denoted by T_c) having category c in the training set. Then, for each page p we computed 600 features (6 times 100), one for each feature Ei ($1 \leq i \leq 6$) and for each category $c \in C$ as follows:

[6]Wikipedia has special categories to group protected pages according to the protection type. We excluded these categories in the computation of the category-based features.

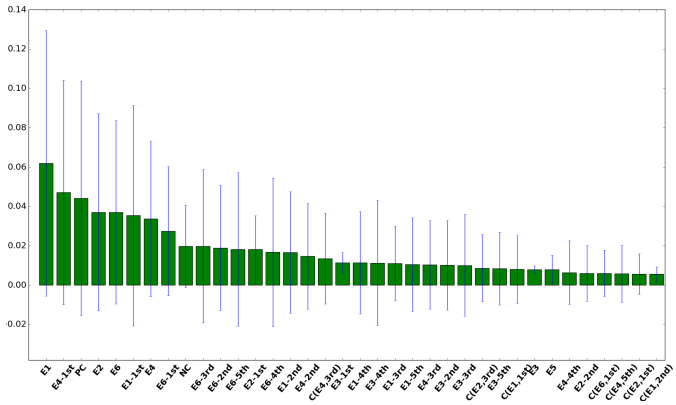

Figure 1: Feature importance (top-36 features).

$$C(Ei, c) = \begin{cases} |Ei(p) - \texttt{avg}_{p' \in T_c}(Ei(p'))| & \text{if } p \text{ is in category } c \\ 0 & \text{otherwise} \end{cases}$$

where $Ei(p)$ is the value of the feature Ei for the page p. The aim of this group of features is to understand if a page is anomalous w.r.t. other pages in the same category.

In summary, in this paper we propose a total of 638 features for DePP, out of which 36 consider the page revision history, 2 consider the page categories only, and 600 consider the variation w.r.t. the page revision features within each category.

5. EXPERIMENTS

We implemented the proposed features and tested their accuracy in the prediction task with 10-fold cross validation by using different classification algorithms, namely logistic regression, support vector machine, k-nearest neighbor, and random forest.

The accuracy achieved by each of these classifiers is shown in Table 1. As we can see, random forest achieves the highest mean accuracy value of 93.24% over 10-folds of the cross validation. This shows that DePP is able to classify pages to protect from pages that do not need protection, very efficiently.

Features Importance. In order to better understand our features, we used forests of 250 randomized trees to determine feature importance. The relative importance (for the classification task) of a feature f in a set of features is given by the depth of f when it is used as a decision node in a tree. Features used at the top of the tree contribute to the final prediction decision of a larger fraction of the input samples. The expected fraction of the samples they contribute to can thus be used as an estimate of the relative importance of the features. Figure 1 shows the importance of the top-36 features for the classification task. The green bars in the plot show the feature importance using the whole forest, while the blue bars represent the variability across the trees. The top-3 most important features are:

- Total average time between revisions (E1): we observe, in our dataset, that the mean average time between revisions is 5.8 days for protected pages and 2.9 months for unprotected ones.

- Total number of revisions by non-registered users in the latest two weeks (E4-1st): in the most recent two weeks, non-registered users make more edits on protected pages (10 edits on average) than on unprotected ones (0.25 edits on average).

- Probability of protecting the page given its categories (PC): a protected page is more likely to be in categories that have other protected pages than an unprotected page (a probability of 0.84 on average vs. 0.52).

Overall, in the top-36 most important features we have all our basic features E1-E6, 9 category-based features, and the majority of the features E1-E6 computed in different time intervals.

Comparison with Baselines. To the best of our knowledge, there is no automated tool in Wikipedia detecting which page to protect, but all the work is manually done by administrators. One of the main reasons for protecting a page on Wikipedia is to stop edit wars, vandalism or libel from happening, or continuing to happen on a page. Thus, we compare DePP with the following baselines:

[B1] *Number of revisions tagged as "Possible libel or vandalism"*: These tags are added automatically without human interference by checking for certain words that might be likely to be vandalism. If a match is found, the tag is added.

[B2] *Number of revisions that Cluebot NG or STiki reverted as possible vandalism*: number of reverted edits in the page made by each one of these two tools.

[B3] *Number of edit wars between two users in the page*: Edit warring occurs when two users do not agree on the content of a page or revision. Therefore, we count the number of edit wars within the revision history of a page as another baseline. We define an edit war as one user making a revision to a page, followed by another user reverting that revision, and this pattern happens 2 or 3 consecutive times.

The performances of the above baselines are shown in Table 1. By considering the best classifier (random forest), baselines based on number of vandalism edits in a page (B1 and B2) present very poor accuracy performances (55.96% and 65.55% resp.). As studied in [5], vandals surf Wikipedia pages link-wise, category wise, or randomly. So, even if they start vandalizing a page that then becomes protected, they can successively move to any other arbitrary page, so B1 and B2 seem not to be good indicators of page protection.

Baseline B3, based on edit wars, performs better than the other two with an accuracy of 73.36%, confirming the fact that at least one edit war happens in pages that end up protected[7]. All baselines together reach an accuracy of 78.09% and are significantly beaten by DePP. By combining DePP with all the three baselines, the accuracy does not improve (93.327% with random forest).

	Baselines
B1	55.964%
B2	65.617%
B3	73.361%
B1+B2+B3	78.089%
	DePP
SVM	89.275%
Logistic Regression	90.816%
K Nearest Neighbor	88.225%
Random Forest	**93.237%**

Table 1: DePP accuracy results and comparison with baselines. Baselines are computed with random forest.

6. CONCLUSIONS

In this paper we proposed DePP, the first system detecting pages to protect in Wikipedia. DePP leverages features based on users page revision behavior and page categories. We built a new dataset to test our system containing 13.6K protected and unprotected pages and 1.9M edits. We showed that DePP performs very well with an accuracy of 93.24% and significantly beats baselines.

As future work we plan to implement and test DePP directly on Wikipedia. DePP does not look at edit content, so it can work with all the different language versions of Wikipedia. Moreover, we would like to extend our data classification task to not just consider which pages require protection versus do not require protection, but attempt to classify the type of protection, including semi-protected or fully protected, and for how long a page my need to be protected, such as a short period of time or indefinitely.

[7]Average number of edit wars in protected pages is 1.37 while the same number for unprotected pages is 0.06.

7. REFERENCES

[1] http://en.wikipedia.org/wiki/User:ClueBot_NG.

[2] http://en.wikipedia.org/wiki/Wikipedia:STiki.

[3] B. T. Adler, L. de Alfaro, S. M. Mola-Velasco, P. Rosso, and A. G. West. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *CICLing*, pages 277–288, 2011.

[4] B. M. Hill and A. D. Shaw. Page protection: another missing dimension of wikipedia research. In *OpenSym*, pages 15:1–15:4, 2015.

[5] S. Kumar, F. Spezzano, and V. S. Subrahmanian. VEWS: A wikipedia vandal early warning system. In *SIGKDD*, pages 607–616, 2015.

[6] S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *WWW*, pages 591–602, 2016.

[7] D. W. McDonald, S. Javanmardi, and M. Zachry. Finding patterns in behavioral observations by automatically labeling forms of wikiwork in barnstars. In *OpenSym*, pages 15–24, 2011.

[8] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in wikipedia. In *GROUP*, pages 259–268, 2007.

[9] H. Roitman, S. Hummel, E. Rabinovich, B. Sznajder, N. Slonim, and E. Aharoni. On the retrieval of wikipedia articles containing claims on controversial topics. In *WWW Companion*, pages 991–996, 2016.

[10] F. B. Viégas, M. Wattenberg, and M. M. McKeon. The hidden order of wikipedia. In *OCSC*, pages 445–454, 2007.