Automatic Vandalism Detection in Wikipedia

Martin Potthast, Benno Stein, and Robert Gerling

Bauhaus University Weimar, Faculty of Media, 99421 Weimar, Germany
<first name>.<last name>@medien.uni-weimar.de

Abstract. We present results of a new approach to detect destructive article revisions, so-called vandalism, in Wikipedia. Vandalism detection is a one-class classification problem, where vandalism edits are the target to be identified among all revisions. Interestingly, vandalism detection has not been addressed in the Information Retrieval literature by now. In this paper we discuss the characteristics of vandalism as humans recognize it and develop features to render vandalism detection as a machine learning task. We compiled a large number of vandalism edits in a corpus, which allows for the comparison of existing and new detection approaches. Using logistic regression we achieve 83% precision at 77% recall with our model. Compared to the rule-based methods that are currently applied in Wikipedia, our approach increases the F-Measure performance by 49% while being faster at the same time.

Introduction. The content of the well-known Web encyclopedia Wikipedia is created collaboratively by volunteers. Every visitor of a Wikipedia Web site can participate immediately in the authoring process: articles are created, edited, or deleted without need for authentication. In practice, an article is developed incrementally since, ideally, authors review and revise the work of others. Till this day about 8 million articles in 253 languages have been authored in this way.

However, all times the Wikipedia and its freedom of editing has been misused by some editors. We distinguish them into three groups: (*i*) lobbyists, who try to push their own agenda, (*ii*) spammers, who solicit products or services, and (*iii*) vandals, who deliberately destroy the work of others. The Wikipedia community has developed policies for a manual recognition and handling of such cases, but enforcing them requires the manpower of many. With the rapid growth of Wikipedia a shift from article contributors to editors working on article maintenance is observed. Hence it is surprising that there is little research to support editors from the latter group or to automatize their tasks. As part of our research Table 1 surveys the existing tools for the prevention of editing misuse.

Related Work. The first attempt to aid lobbying detection was the WikiScanner tool which maps IP numbers recorded from anonymous editors to their domain name. This way editors can be found who are biased with respect to the topic in question. Since there are diverse ways for lobbyists to disguise their identity a manual check of all edits for hints of lobbying is still necessary.

There has been much research concerning spam detection in e-mails, among Web pages, or in blogs. In general, machine learning approaches, possibly combined with

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 663–668, 2008.

Table 1. Tools for the prevention of editing misuse with respect to the target group, and the type of automation (aid, full). Tools shown gray use the same or a very similar rule set as the tool listed in the line above.

Tool	Target	Type	Status	URL (October 2007)
WikiScanner	lobbyists	aid	active	http://wikiscanner.virgil.gr
AntiVandalBot (AVB)	vandals	full	inactive	http://en.wikipedia.org/wiki/WP:AVB
MartinBot	vandals	full	inactive	http://en.wikipedia.org/wiki/User:MartinBot
T-850 Robotic Assistant	vandals	full	active	http://en.wikipedia.org/wiki/User:T-850_Robotic_Assistant
WerdnaAntiVandalBot	vandals	full	active	http://en.wikipedia.org/wiki/User:WerdnaAntiVandalBot
Xenophon	vandals	full	active	http://en.wikipedia.org/wiki/User:Xenophon_(bot)
ClueBot	vandals	full	active	http://en.wikipedia.org/wiki/User:ClueBot
CounterVandalismBot	vandals	full	active	http://en.wikipedia.org/wiki/User:CounterVandalismBot
PkgBot	vandals	aid	active	http://meta.wikimedia.org/wiki/CVN/Bots
MiszaBot	vandals	aid	active	http://en.wikipedia.org/wiki/User:MiszaBot

manually developed rules, do an excellent spam detection job [1]. The respective technology may also be adequate for a misuse analysis in Wikipedia, but the applicability has not been investigated yet.

Vandalism was recognized as an open problem by researchers studying online collaboration [2,4,5,6,7,8], and, of course, by the Wikipedia community.[1] The former provide statistical or empirical analyses concerning vandalism, but neglect its detection. The latter developed four small sets of detection rules but did not evaluate the performance. Misuses such as trolling and flame wars in discussion boards are related to vandalism, but so far no research exists to detect either of them.

In this paper we develop foundations for an automatic vandalism detection in Wikipedia: (*i*) we define vandalism detection as a classification task, (*ii*) discuss the characteristics by which humans recognize vandalism, and (*iii*) develop tailored features to quantify them. (*iv*) A machine-readable corpus of vandalism edits is provided as a common baseline for future research. (*v*) Finally, we report on experiments related to vandalism detection based on this corpus.

Vandalism Detection Task. Let $E = \{e_1, \ldots, e_n\}$ denote a set of edits, where each edit e comprises two consecutive revisions of the same document d from Wikipedia, say, $e = (d_t, d_{t+1})$. Let $\mathcal{F} = \{f_1, \ldots, f_p\}$ denote a set of vandalism indicating features where each feature f_i is a function that maps edits onto real numbers, $f_i : E \rightarrow \mathbf{R}$. Using $\mathcal{F}$ an edit e is represented as a vector $\mathbf{e} = (f_1(e), \ldots, f_p(e))$; $\mathbf{E}$ is the set of edit representations for the edits in E.

Given a vandalism corpus E which has a realistic ratio of edits classified as vandalism and well-intentioned edits, a classifier c, $c : \mathbf{E} \rightarrow \{0, 1\}$, is trained with examples from E. c serves as an approximation of c^*, the true predictor of the fact whether or not an edit forms a vandalism case. Using $\mathcal{F}$ and c one can classify an edit e as vandalism by computing $c(\mathbf{e})$.

[1] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Vandalism_studies (October 2007)

Table 2. Organization of vandalism edits along the dimensions "Edited content" and "Editing category": the matrix shows for each combination the portion of specific vandalism edits at all vandalism edits. For vandalized structure insertion edits and content insertion edits also a list of their typical characteristics is given. It includes both the characteristics described in the previous research and the Wikipedia policies.

Editing category	**Edited content**			
	Text	Structure	Link	Media
Insertion	43.9% Characteristics: point of view, off topic, nonsense, vulgarism, duplication, gobbledegook	14.6% Characteristics: formatting, highlighting	6.9%	0.7%
Replacement	45.8%	15.5%	4.7%	2.0%
Deletion	31.6%	20.3%	22.9%	19.4%

Vandalism Indicating Features. We have manually analyzed 301 cases of vandalism to learn about their characteristics and, based on these insights, to develop a feature set $\mathcal{F}$. Table 2 organizes our findings as a matrix of vandalism edits along the dimensions "Edited content" and "Editing category"; Table 3 summarizes our features.

Table 3. Features which quantify the characteristics of vandalism in Wikipedia

Feature f	**Description**
char distribution	deviation of the edit's character distribution from the expectation
char sequence	longest consecutive sequence of the same character in an edit
compressibility	compression rate of an edit's text
upper case ratio	ratio of upper case letters to all letters of an edit's text
term frequency	average relative frequency of an edit's words in the new revision
longest word	length of the longest word
pronoun frequency	number of pronouns relative to the number of an edit's words (only first-person and second-person pronouns are considered)
pronoun impact	percentage by which an edit's pronouns increase the number of pronouns in the new revision
vulgarism frequency	number of vulgar words relative to the number of an edit's words
vulgarism impact	percentage by which an edit's vulgar words increase the number of vulgar words in the new revision
size ratio	the size of the new version compared to the size of the old one
replacement similarity	similarity of deleted text to the text inserted in exchange
context relation	similarity of the new version to Wikipedia articles found for keywords extracted from the inserted text
anonymity	whether an edit was submitted anonymously, or not
comment length	the character length of the comment supplied with an edit
edits per user	number of previously submitted edits from the same editor or IP

For two vandalism categories the matrix shows particular characteristics by which an edit is recognized as vandalism: a vandalism edit has the "point of view" characteristic if the vandal expresses personal opinion, which often entails the use of personal pronouns. Many vandalism edits introduce off-topic text with respect to the surrounding text, are nonsense in that they contradict common sense, or do not form a correct sentence from their language. The first three characteristics are very difficult to be quantified, and research in this direction will be necessary to develop reliable analysis methods. Vulgar vandalism can be detected with a dictionary of vulgar words; however, one has to consider the context of a vulgar word since several Wikipedia articles contain vulgar words in a correct sense. Hence we quantify the impact of a vulgar word based on the point of time it has been inserted into an article rather than simply checking its occurrence. If an inserted text duplicates other text within the article or within Wikipedia, one may also speak of vandalism, but this is presumably the least offending case. Very often vandalism consists only of gobbledygook: a string of characters which has no meaning whatsoever, for instance if the keyboard is hit randomly. Another common characteristic of vandalism is that it is often highlighted by capital letters or by the repetition of characters. In cases of deletion vandalism, larger parts of an article are deleted, which explains the high percentages of this vandalism type throughout all content types. Note that a vandalism edit typically shows several of these characteristics at the same time.

Vandalism Corpus. Vandalism is currently not documented in Wikipedia, so that automatic vandalism detection algorithms cannot be compared to each other. The best way to find vandalism manually is by taking a look at the list of the most vandalized pages and then to analyze the history of the listed articles.[2] We have set up the vandalism corpus WEBIS-VC07-11, which was compiled from our own investigations and the results of a study[3] conducted by editors of Wikipedia. The corpus contains 940 human-assessed edits from which 301 edits are classified as vandalism. It is available in a machine-readable form for download at [9].

Evaluation. Within one-class classification tasks one is often confronted with the problem of class imbalance: one of the classes, either the target or the outlier class is underrepresented, which makes training a classifier difficult. In a realistic detection scenario only 5% of all edits in a given time period are from the target class "vandalism" [5]. As a heuristic to alleviate the problem we resort to random over-sampling of the underrepresented class at training time. Nevertheless, an in-depth analysis with respect to domain characteristics of the training samples is still necessary; the authors of [3] have compared alternative methods to address class imbalance.

Using ten-fold cross-validation on the corpus WEBIS-VC07-11 and a classifier based on logistic regression we evaluated the discriminative power of the features described in Table 3 when telling apart vandalism and well-intentioned edits. We also analyzed the effort for computing these features and compared the results to AVB and to ClueBot. Table 4 summarizes the results.

As can be seen, our approach (third row) outperforms the rule-based bots on all accounts. The individual analysis of each feature indicates its contribution to the overall

[2] http://en.wikipedia.org/wiki/Wikipedia:Most_vandalized_pages (October 2007)
[3] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Vandalism_studies/Study1 (Oct. 2007)

Table 4. Vandalism detection performance quantified as category-specific recall and averaged precision values. The first row shows, as the baseline, the currently best performing Wikipedia bot, while the third row (bold) shows the results of our classifier. The right column shows the throughput on a standard PC. The underlying test corpus contains 940 human-assessed edits from which 301 edits are classified as vandalism.

Feature f	**Recall**			**Precision**	**Throughput**
	Insertion	Replacement	Deletion	Average	(edits per second)
Baseline: AVB	0.35	0.53	0.61	0.74	3
ClueBot	0.03	0.29	0.49	1	3
c **with all features**	**0.87**	**0.76**	**0.89**	**0.86**	**5**
char distribution	0.03	0	0.74	0.41	6
char sequence	0.01	0.14	0.2	0.70	43
compressibility	0	0	0.78	0.24	618
upper case ratio	0.13	0.22	0	0.61	656
term frequency	0	0.29	0.01	0.3	4
longest word	0	0.04	0.63	0.54	319
pronoun frequency	0.09	0.1	0	0.53	351
pronoun impact	0	0.04	0.39	0.49	53
vulgarism frequency	0.23	0.35	0	0.65	181
vulgarism impact	0.23	0.41	0.52	0.91	33
size ratio	0.07	0.35	0.54	0.83	8198
replacement similarity	–	0	–	–	9
context relation	0	0	0.13	0.18	3
anonymity	0	0	0	0	8 545
comment length	0	0	0	0	14 242
edits per user	0.94	0.86	0.96	0.66	813

performance. Note that vandalism detection suggests a two-stage analysis process (machine + human) and hence to prefer high recall over high precision: a manual post-processing of classifier results is indispensable since visitors of a Wikipedia page should never see a vandalized document; as well as that, a manual analysis is feasible because an even imprecisely retrieved target class contains only few elements.

References

1. Blanzieri, E., Bryl, A.: A Survey of Anti-Spam Techniques. Technical Report DIT-06-056, University of Trento (2006)
2. Buriol, L.S., Castillo, C., Donato, D., Leonardi, S., Millozzi, S.: Temporal Analysis of the Wikigraph. In: WI 2006, pp. 45–51. IEEE Computer Society, Los Alamitos (2006)
3. Japkowicz, N., Stephen, S.: The Class Imbalance Problem: A Systematic Study. Intell. Data Anal. 6(5), 429–449 (2002)
4. Kittur, A., Suh, B., Pendleton, B., Chi, E.: He says, she says: Conflict and Coordination in Wikipedia. In: CHI 2007, pp. 453–462. ACM, New York (2007)

5. Priedhorsky, R., Chen, J., Lam, S., Panciera, K., Terveen, L., Riedl, J.: Creating, Destroying, and Restoring Value in Wikipedia. In: Group 2007 (2007)
6. Viégas, F.B.: The Visual Side of Wikipedia. In: HICSS 2007, p. 85. IEEE Computer Society, Los Alamitos (2007)
7. Viégas, F.B., Wattenberg, M., Dave, K.: Studying Cooperation and Conflict between Authors with History Flow Visualizations. In: CHI 2004, pp. 575–582. ACM Press, New York (2004)
8. Viégas, F.B., Wattenberg, M., Kriss, J., van Ham, F.: Talk before you Type: Coordination in Wikipedia. In: HICSS 2007, p. 78. IEEE Computer Society, Los Alamitos (2007)
9. Potthast, M., Gerling, R. (eds): Web Technology & Information Systems Group, Bauhaus University Weimar. Wikipedia Vandalism Corpus WEBIS-VC07-11 (2007), http://www.uni-weimar.de/medien/webis/research/corpora