ClueBot and Vandalism on Wikipedia

by

Naomi Persephone Amethyst

Introduction

Imagine you have just spent three days completing an article for Wikipedia, submitted it, viewed it and all is well. The next day you are at your friend's house, and you say, “Hey, check out this article I created!” When you open it, however, your article is filled with profanities! Imagine your disappointment and anger. Vandalism occurs on Wikipedia far too often. Something needs to be done about it.

Wikipedia, *The Free Encyclopedia,* is an online encyclopedia which is editable by anyone. This affords readers the unique opportunity to add their knowledge to the encyclopedia, to correct mistakes and typographical errors, and to change the wording such that the articles read better. It is by no means mandatory to contribute, but people do it voluntarily out of good will and loyalty to the project. Indeed, it is very easy to edit the encyclopedia. All one has to do is click the edit button and proceed to change the page content. One doesn't even have to sign up to edit, although it is encouraged to do so. In a perfect world, all of Wikipedia's users would make helpful contributions in good faith. Unfortunately, such a utopia does not exist. Not all users have the community's best interests at heart while editing, and not all edits are good. Some users even have malicious intent, and deliberately try to destroy all the hard work that has gone into Wikipedia by removing all of the content from a page or replacing the page with obscenities. These types of edits are considered online vandalism. Such vandals have a twisted sense of humor. They think that destroying others' hard work is fun.

Many pages get quite a lot of vandalism. Because of this, many dedicated editors who would otherwise be writing articles and improving existing articles are instead cleaning up after these malicious vandals. There exist many tools to fight such vandalism, but most require the constant

attention of a human operator. Most such tools are simply add-ons to the users' web browser which facilitate the manual removal of vandalism. Fixing vandalism takes up a lot of users' time on Wikipedia, and it would be very useful to have a machine fix these problems instead of manual human intervention.

Wikipedia has a system where old versions of articles are stored and accessible by users, no matter what happens to the current version. This revision system has kept track of all the changes to all articles since Wikipedia's inception. This allows users to see the state of an article at any point in time and allows users to revert a page back to an older version of the page if it is changed in a malicious way.

Humans can indeed fight vandalism, but it is a very inefficient use of time. I thought to myself: "Why can't a computer fix vandalism all by itself?" Having had quite a bit of experience with automated systems, I decided to create an automatic method to fix vandalism.

There have been attempts at this before. MartinBot, for example, is a computer program which is designed to find and revert vandalism. Unfortunately, MartinBot is often offline, and even when it is running, it makes mistakes such as reverting legitimate edits (false positives) and not detecting

real vandalism. MartinBot also uses a lot of resources on the toolserver (the official Wikipedia server on which it runs). According to one Wikipedia administrator, MartinBot is the single most resource-intensive program that is run on the toolserver.

I decided that I should write a program that does what MartinBot does but fixes MartinBot's problems. The goal was to make a program that could efficiently and reliably detect and correct instances of vandalism and simultaneously fix the resource consumption problem by moving the vandalism detection routines to an external server outside of the official Wikipedia network.

This new program is called ClueBot. The basic structure of the program is simple. It constantly checks Wikipedia for new page edits by users. When it detects a new edit, it analyzes it. If it determines that the edit is vandalism, then it fixes the vandalism by reverting the article to its previous state. It also takes some steps to ensure that vandalism by the same user won't happen again, in accordance with Wikipedia policy. [1] Once ClueBot finishes with a given instance of vandalism, it returns to checking for new edits. [2]

The first thing I had to do was write the parts of the program to interface with Wikipedia. This was essentially routine programming. Once I finished

1 http://en.wikipedia.org/wiki/WP:VAN#Dealing_with_vandalism
2 See the flow chart attached to the end of this report.

writing these functions, I started working on vandalism detection. A human can easily detect vandalism just by glancing at a page. However, to a computer, it's all just text. A computer must be explicitly told what to count as vandalism and what to ignore. This can be very difficult because there are many diverse types of vandalism and many types are even subjectively determined.

The set of rules followed to try to detect vandalism are called heuristics. I noticed several similarities between many types of malicious edits which I could use in the detection portion of the program. Such edits often did one of the following: removed all content from the page, replaced legitimate page content with worthless junk, removed a massive amount of the content, or added a massive amount of nonsense. Because of these similarities, I created four core rules[3]. Note that these four core rules alone are not very effective at determining vandalism. To rectify this, I created a scoring system which scores an edit based on what is added and removed. Scoring rules are applied to a point total for each edit. Once all of the scoring rules have been applied, a total number of points is generated. This is the edit's score.

The first of the four core rules tested is a rule that checks whether or not an edit replaces the page's entire content with something else. This test is triggered most frequently because most vandals simply replace the entire page

3 See chart attached to the end of this report.

with what they want.

The second core rule tested checks whether or not an edit removes all the page's content. This is called blanking the page. This test is triggered the second most frequently because a lot of vandals can't be bothered to even generate their own content.

The third core rule tests if an edit removes more than 10,000 characters from the page. This rule required substantial amounts of modification to settle on a constant number of characters. At first, 10,000 characters was a bit high, so I changed it to 5,000 characters. This ended up producing many false positives. I finally settled on 7,500 characters which seems to be a good number with very few errors. Furthermore, this rule checks to see if the score of the edit is less than -50 points before being applied.

The fourth and final core rule checks if an edit adds 7,500 characters and the edit's current score is less than -1,000 points. This rule triggers when someone adds nonsense or obscenities to the page.

An edit's score is determined for a given edit by starting the point total at zero. If an edit adds obscenities, points are removed. If it removes obscenities, points are added. Edits that look like personal attacks cause

points to be lost, and edits that remove personal attacks cause points to be gained. Edits that add bad grammar or remove good grammar cause points to be removed, and edits that add good grammar or remove bad grammar cause points to be added. Removal of information boxes or other helpful Wikipedia-specific content will also cause a deduction of points, whereas addition of such content will add points.

I also noticed that established users rarely vandalized pages, but anonymous users or newly created users often did. So, I added a rule such that any edit by a user with more than 50 edits is not considered to be vandalism. Likewise, any edit by an anonymous user who has made more than 250 edits is defined to not be vandalism, either. An additional user-based rule that I added as a failsafe is a “whitelist” - a list of users which should never be counted as making vandalism.

Because it is possible for ClueBot to come to a faulty conclusion about an edit, it does not revert back to its own previous revisions, nor will it revert the same user and page combination twice in the same 24 hours. This makes it such that if ClueBot is wrong, the user can just redo what they did and ClueBot will let them do it. Vandals will rarely vandalize the same page twice in a row, because they usually act on impulse.

Once ClueBot was sufficiently advanced to reliably detect vandalism, it was submitted for approval to Wikipedia's Bot Approval Group. The BAG is responsible for approving which automated programs are allowed to make edits to Wikipedia. After a bit of discussion and several good suggestions (which were implemented), ClueBot was approved for a fifty-revert trial, meaning that it could only fix fifty instances of vandalism. This first trial had some stipulations: the first forty reverts must be checked manually before allowing the bot to fix the vandalism. The last ten could be done automatically by the bot without intervention. The trial went very smoothly and it was shortly approved for a full fortnight trial. The second trial was to be fully automatic and didn't require a human operator to watch each revert, as the bot was to run continuously — with the exception of program edits — for the duration of the two week trial. That trial went so well that ClueBot was approved for continuous operation two days before the fortnight trial was scheduled to end.

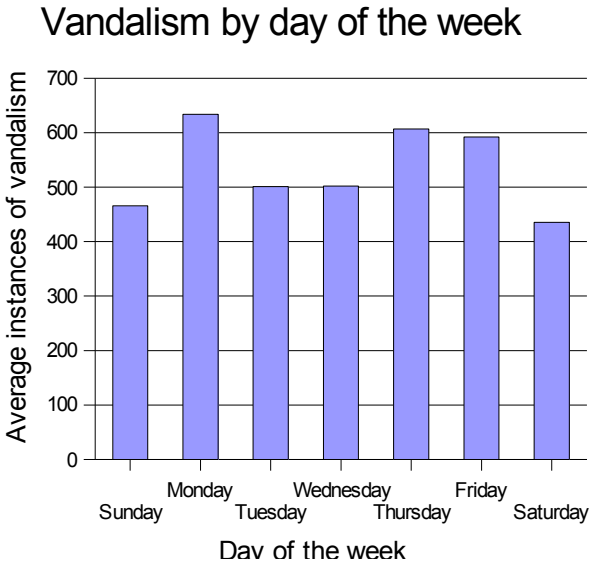
From July 23rd, 2007 to September 25th, 2007, ClueBot has corrected over 21,000 instances of vandalism[4]. Its high efficiency, reliability, and helpfulness have caused it to gain a large amount of praise. On Wikipedia, there is a system by which the community can hand out awards to specific people or automated programs for exceptional contributions to Wikipedia. These awards are called Barnstars. Cluebot has received 14 Barnstars since its

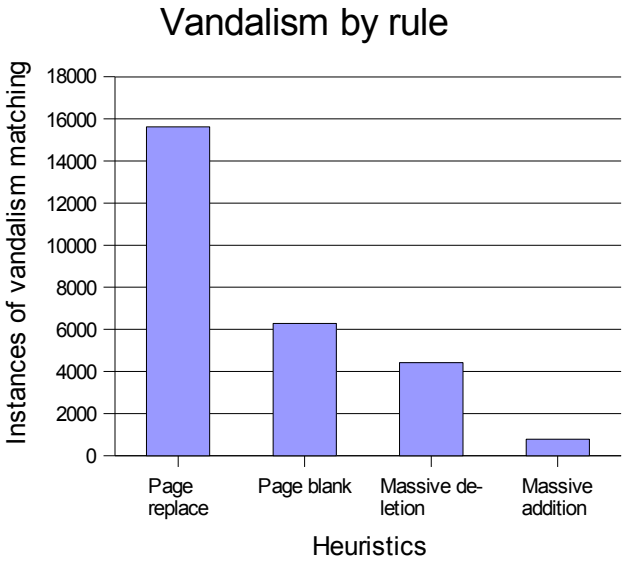
4 See charts attached to the end of this report.

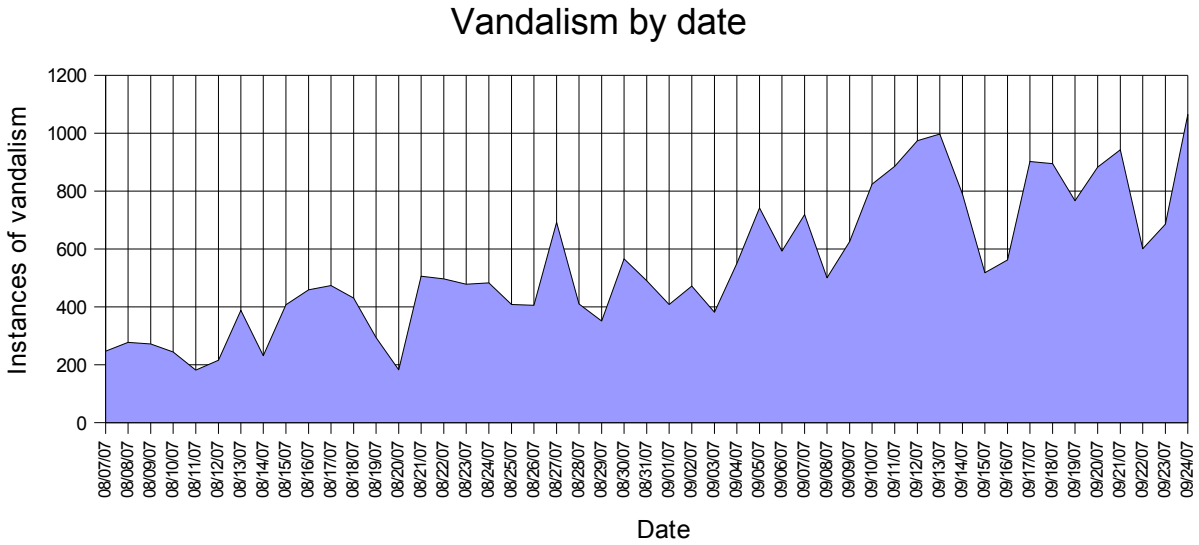
inception. ClueBot has even inspired other users to pursue their own interests in creating automated vandalism protection systems.

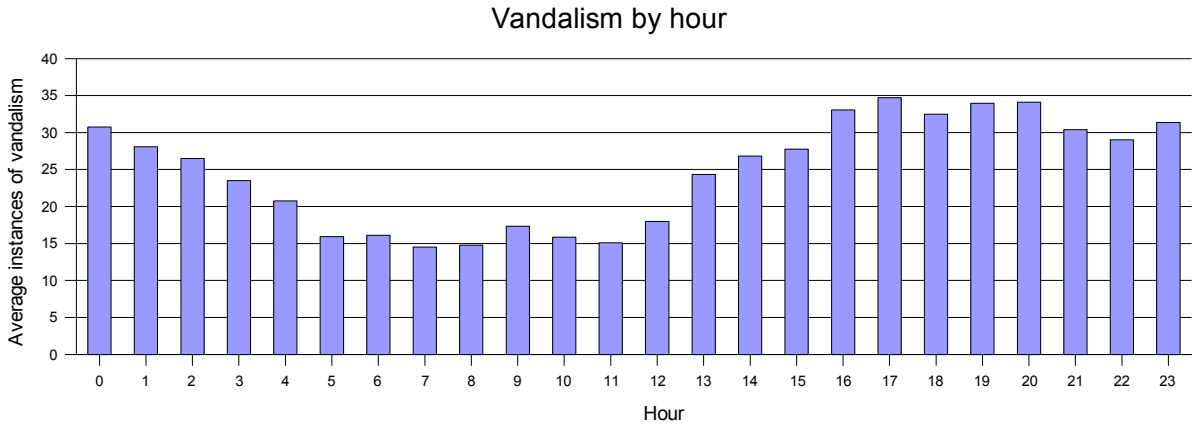
Overall, ClueBot has been a very worthwhile project. I have spent numerous hours on it and have written countless lines of programming code. The end product is a finely tuned and very useful piece of software. It has extensively benefited the community as shown by its many awards and commendations, and it has helped facilitate informational exchange over the Internet.

In addition to its immediate benefits, the methods employed in the identification of vandalism have helped extend the frontier of automatic interpretation of arbitrary human data, a rapidly expanding field. The development of ClueBot has been a rewarding experience for me, other programmers and researchers, and the community at large. I hope it will continue to run for years to come, and I will continue to update it as new methods are invented to automatically distinguish useful data from malicious edits.

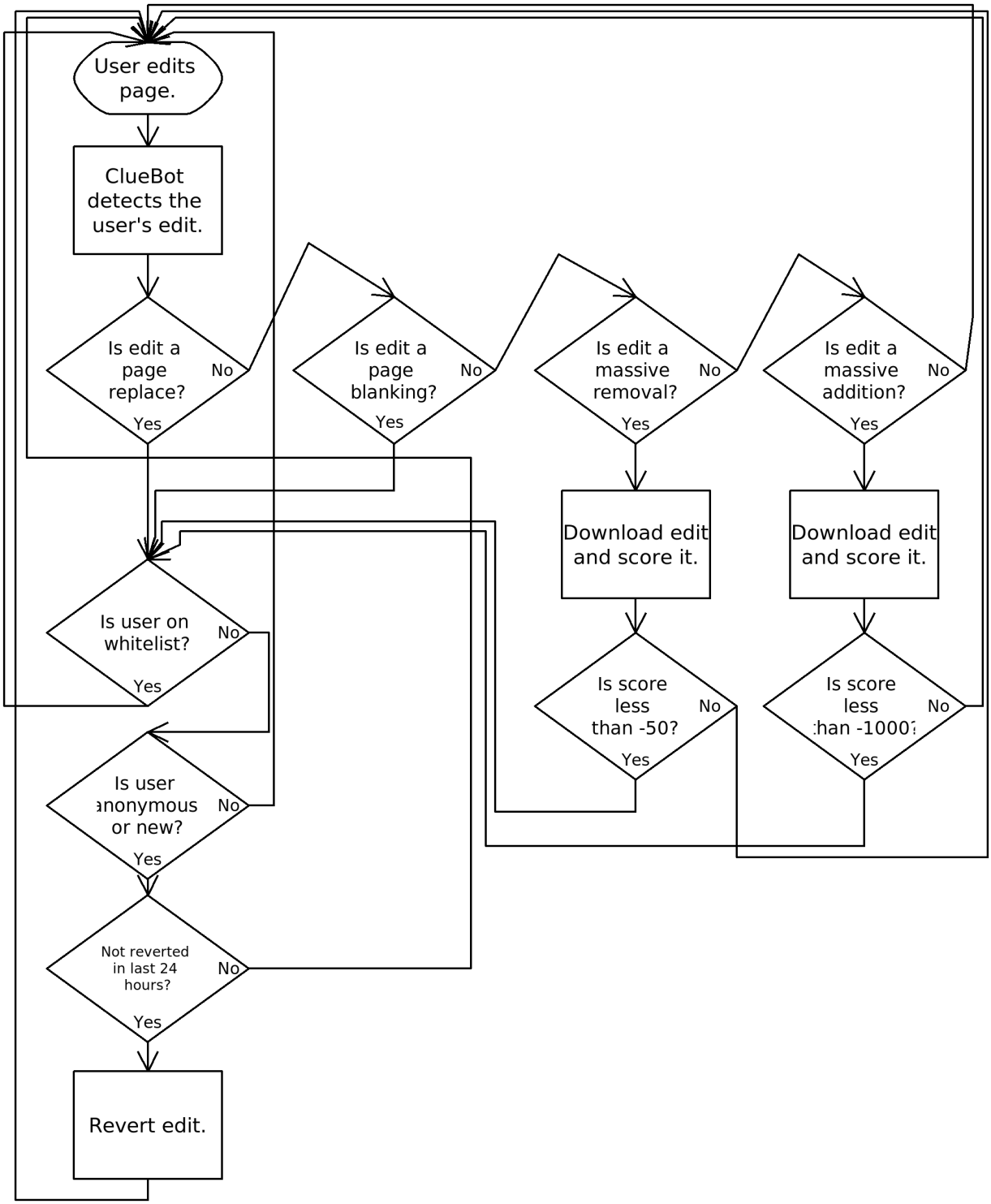

Vandalism by day of the week
Average instances of vandalism
700
600
500
400
300
200
100
0
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of the week


Vandalism by rule
Instances of vandalism matching
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
Page replace
Page blank
Massive de-letion
Massive addition
Heuristics


Vandalism by date
Instances of vandalism
1200
1000
800
600
400
200
0
08/07/07
08/08/07
08/09/07
08/10/07
08/11/07
08/12/07
08/13/07
08/14/07
08/15/07
08/16/07
08/17/07
08/18/07
08/19/07
08/20/07
08/21/07
08/22/07
08/23/07
08/24/07
08/25/07
08/26/07
08/27/07
08/28/07
08/29/07
08/30/07
08/31/07
09/01/07
09/02/07
09/03/07
09/04/07
09/05/07
09/06/07
09/07/07
09/08/07
09/09/07
09/10/07
09/11/07
09/12/07
09/13/07
09/14/07
09/15/07
09/16/07
09/17/07
09/18/07
09/19/07
09/20/07
09/21/07
09/22/07
09/23/07
09/24/07
Date


Vandalism by hour
Average instances of vandalism
40
35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour

Heuristics flow chart



References

"Wikipedia's Bot Policy." *Wikipedia, the Free Encyclopedia*. 03:41, 21 September 2007 UTC. Wikimedia Foundation. 25 Sep. 2007 <http://en.wikipedia.org/wiki/WP:BOT>.

"Wikipedia's Bot Approval Group." *Wikipedia, the Free Encyclopedia*. 21:34, 24 September 2007 UTC. Wikimedia Foundation. 25 Sep. 2007 <http://en.wikipedia.org/wiki/WP:BAG>.

"Dealing with vandalism." *Wikipedia, the Free Encyclopedia*. 06:21, 24 September 2007 UTC. Wikimedia Foundation. 25 Sep. 2007 <http://en.wikipedia.org/wiki/WP:VAN#Dealing_with_vandalism>.

"MartinBot." *Wikipedia, the Free Encyclopedia*. 01:03, 24 September 2007 UTC. Wikimedia Foundation. 25 Sep. 2007 <http://en.wikipedia.org/wiki/User:MartinBot>.

"ClueBot." *Wikipedia, the Free Encyclopedia*. 19:42, 13 September 2007 UTC. Wikimedia Foundation. 25 Sep. 2007 <http://en.wikipedia.org/wiki/User:ClueBot>.

"Barnstars." *Wikipedia, the Free Encyclopedia*. 21:50, 17 September 2007 UTC. Wikimedia Foundation. 25 Sep. 2007 <http://en.wikipedia.org/wiki/WP:BARN>.