Bots and Cyborgs: Wikipedia's Immune System

Aaron Halfaker and John Riedl
University of Minnesota

Bots and cyborgs are more than tools to better manage content quality on Wikipedia—through their interaction with humans, they're fundamentally changing its culture.

When Wikipedia was young and the number of active contributors numbered in the 10s or 100s, volunteer editors could directly manage its content and processes. Some editors who have been around the longest will fondly recall the halcyon days when the encyclopedia evolved slowly, and one person could manually track all of a day's changes in a matter of minutes.

Those days ended in 2004 when Wikipedia began to experience exponential growth in new contributors, new articles, and popular media attention. By the time growth peaked in 2007, the encyclopedia was receiving more than 180 edits per minute.

On one hand, Wikipedia was enjoying a rich flow of its lifeblood: volunteer contributions. On the other hand, suddenly no human could review all of the changes. Experienced editors spent all of their time watching for copyright violations, potentially libelous articles, and vandalism; they simply couldn't keep up with the volume of incoming edits.

Community members responded to this problem by developing two basic types of computational tools: robots, or *bots,* and *cyborgs*. Bots automatically handle repetitive tasks—for example, SpellCheckerBot fixes spelling errors. The sidebar "A Taxonomy of Wikipedia Bots" provides an overview of the many activities bots perform on the site. Cyborgs are intelligent user interfaces that humans "put on" like a virtual Ironman suit to combine computational power with human reasoning. Huggle, for instance, helps users zap vandals' edits by the thousands.

Together, bots and cyborgs function as first-line defenses in Wikipedia's emerging "immune system."

COMBATING VANDALISM

Bots and cyborgs arguably have been most effective at filtering vandalism on Wikipedia. Along with the 2004-2007 exponential growth in good contributions came a torrent of bad content and damaging edits that humans alone couldn't handle.

Early tools

The first tools to redefine the way Wikipedia dealt with vandalism were AntiVandalBot and VandalProof.

AntiVandalBot used a simple set of rules and heuristics to monitor changes made to articles, identify the most obvious cases of vandalism, and automatically revert them. Although this bot made it possible, for the first time, for the Wikipedia community to protect the encyclopedia from damage without wasting the time and energy of good-faith editors, it wasn't very intelligent and could only correct the most egregious instances of vandalism.

VandalProof, an early cyborg technology, was a graphical user interface written in Visual Basic that let trusted editors monitor article edits as fast as they happened in Wikipedia and revert unwanted contributions in one click. VandalProof served as a natural

0018-9162/12/$31.00 © 2012 IEEE Published by the IEEE Computer Society

A TAXONOMY OF WIKIPEDIA BOTS

Robots perform a wide range of activities on Wikipedia. These include injecting public domain data, monitoring and curating content, augmenting the MediaWiki software, and protecting the encyclopedia from malicious activity.

The first bots injected data into Wikipedia content from public databases. Rambot, widely accepted to be the encyclopedia's first sanctioned robot, inserted census data into articles about countries and cities. Rambot and its cousins act as "force multipliers" by performing repetitive activities hundreds or thousands of times in minutes.

Many other bots monitor and curate Wikipedia content. For example, SpellCheckerBot checks recent changes for common spelling mistakes using an international dictionary to prevent accidental "fixes" to correctly spelled foreign words. Similarly, Helpful Pixie Bot corrects ISBNs and other structural features of articles such as section capitalization. The largest class of content curators is interlanguage bots, which use graph models of links between different languages of Wikipedia to identify missing links between articles covering the same topic in a different language. As of this writing, the English Wikipedia has more than 60 active interlanguage bots.

Some bots extend Wikipedia functionality by implementing features that the MediaWiki software doesn't support. For example, AIV Helperbot turns a simple page into a dynamic priority-based discussion queue to support administrators in their work of identifying and blocking vandals. Similarly, SineBot ensures that every posted comment is signed and dated.

Finally, a series of bots protects the encyclopedia from malicious activity. For example, ClueBot_NG uses state-of-the-art machine learning techniques to review all contributions to articles and to revert vandalism, while XLinkBot reverts contributions that create links to blacklisted domains as a way of quickly and permanently dealing with spammers.

supplement to AntiVandalBot and its successors by leaving the obvious vandalism to bots' simple decision-making algorithms and letting human editors handle the rest.

Current tools

Several years and many iterations later, bots and cyborgs have become more powerful, accurate, and user-friendly. Consequently, as Figure 1 shows, these tools have had an increasing role in maintaining article quality on Wikipedia.

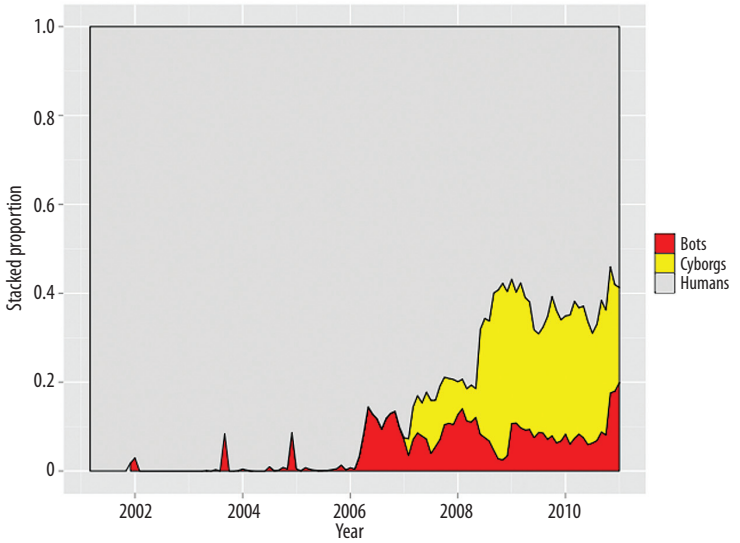

Figure 1. Stacked proportion of reverts (rejections of damaging edits) for bots, cyborgs, and human editors on Wikipedia. Bots and cyborgs have been taking over an increasing role in protecting articles from damage since early 2006.

ClueBot_NG has replaced AntiVandalBot's simple rules and heuristics with a highly accurate neural network/machine learning approach: editors submit examples of mistakes that ClueBot_NG makes, and the cyborg's developers retrain its classifier periodically. Similarly, Huggle has replaced VandalProof with a slick user interface, configurablity, and an intelligent system for sorting edits by vandalistic likelihood to maximize the efficiency of human effort in dealing with those instances of vandalism that ClueBot_NG doesn't catch.

Distributed cognition

Today, the combined efforts of ClueBot_NG and a small group of Huggle-human cyborgs identify and immediately destroy the majority of vandalism before any human editor or reader ever sees it. However, the speed and efficiency with which these tools deal with individual instances of vandalism is only one aspect of Wikipedia's antivandal strategy.

R. Stuart Geiger and David Ribes observed that bots and cyborgs form a "distributed cognition system" that has reshaped Wikipedia's process for identifying and removing vandals ("The Work of Sustaining Order in Wikipedia: The Banning of a Vandal," *Proc. 2010 ACM Conf. Computer Supported Cooperative Work* [CSCW 10], ACM, 2010, pp. 117-126).

Although bots and cyborg-editors work independently, they "operationalize each offending edit into a social structure through which administrators and editors come to know users as vandals." In other words, Wikipedia's bots and cyborgs automatically build a record of new editors' activities to quickly and efficiently identify vandals and block them, thereby conferring immunity to future damage from that source.

BOT SOCIALIZATION

While bots are highly productive, performing edits hundreds of times faster than humans, they can also be massively disruptive to the community if they perform inappropriate actions, either because of disagreements between bot authors and other Wikipedians or because of bugs. To control for such issues, the Bot Approvals Group, a band of volunteer members, vets new bot proposals and addresses bot-related grievances.

Geiger argues that Wikipedia editors view bots not only as tools or "force multipliers" but also as social agents ("The Lives of Bots," *Critical Point of View: A Wikipedia Reader,* G. Lovink and N. Tkacz, eds., Institute of Network Cultures, 2011, pp. 78-93).

This is at least partially due to the fact that bots interact with Wikipedia in largely the same manner as human editors: they edit articles and other pages via user accounts, and they send and receive messages, usually read by their human maintainer—at least for now. Tawker, AntiVandalBot's operator, even half-jokingly nominated his bot for election to the 2006 Arbitration Committee (Wikipedia's version of the Supreme Court).

As an example, Geiger describes the intense reaction within the Wikipedia community to HagermanBot, which enforced the commonly accepted guideline of signing comments to ensure that all participants in a conversation were named. Although the bot sometimes mistakenly signed contributions not intended to be comments, its normal approved activities were what caused offense. In the words of one editor, Sensemaker, "I don't really like this bot editing messages on other people's talk pages without either their consent or even knowledge."

Although human editors had been performing HagermanBot's function long before it was born, many wanted to hold the bot to a higher standard.

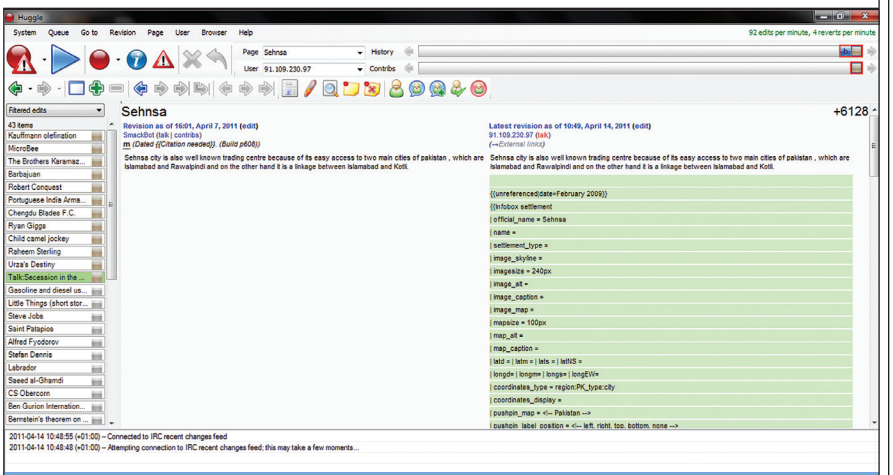

Figure 2. The Huggle interface makes it easy to review a series of recent revisions by filtering them according to the user's preferences.

In the discussions about whether HagermanBot should be allowed to continue to operate, prolific editor and administrator Rich Farmbrough argued it was important for users to understand that "bots are better behaved than people," and cautioned against "botophobia" on Wikipedia. Bots were to some extent becoming social agents, and they would need to find a way to peacefully coexist beside flesh-and-blood contributors.

Luckily for HagermanBot and its operator, an opt-out mechanism settled the issue. Editors like Sensemaker could easily tell the bot not to sign any of their edits by broadcasting their preferences on their profile, which the bot would check before signing a comment. In time, this opt-out mechanism became an operationalization of Isaac Asimov's second law of robotics: a robot must obey the orders given to it by human beings.

MOTIVATIONAL ASPECTS OF CYBORGS

Huggle, one of the most popular antivandalism editing tools on Wikipedia, is written in C#.NET and any user can download and install it. Huggle lets editors *roll back* changes with a single mouse click, but because the tool is so powerful, rollback permission is restricted to administrators and a few thousand other Wikipedia users.

Huggle makes it easy to review a series of recent revisions by filtering them according to the user's preferences. Figure 2 shows the main interface. On the left side is a list of recent revisions sorted by vandalistic probability. When a user selects one of these revisions, it vanishes from the list for all other Huggle users to reduce the risk of conflict. On the right side of the screen, changes made in the selected revision are highlighted, with green indicating added content and yellow deleted content. The far-right number immediately above the revision (+6128 in this example) is the number of words added (or deleted, if the number is negative) in this revision.

The upper right of the interface displays the rate at which the user is working—in this case, the user is reviewing 92 edits per minute and making 4 reverts per minute. In the absence of information about the quality of users' editorial efforts, this "score" encourages frequent reverts. So does the ease of reversion: when the user clicks the large red button with an overlapping exclamation point in the upper left of the interface, Huggle reverts the edit being displayed and sends a warning to the

suspected vandal for inappropriate behavior.

Some Wikipedians feel that such motivational measures have gone too far in making Wikipedia like a game rather than a serious project. One humorous entry even argues that Wikipedia has become a MMORPG—a massively multiplayer online role-playing game—with "monsters" (vandals) to slay, "experience" (edit or revert count) to earn, and "overlords" (administrators) to submit to (http://en.wikipedia.org/wiki/Wikipedia:MMORPG).

Bots and cyborgs have become essential to the Wikipedia ecosystem. Without them, the exponential influx of new users from 2001 to 2007 would never have been manageable. However, bots and cyborgs are more than tools to better manage content quality—through their interaction with humans, they're fundamentally changing Wikipedia's culture.

For instance, recent research demonstrates that reverts are a powerful demotivator to Wikipedia contributors, especially newcomers. Tools like Huggle that automatically identify potential revert-worthy edits, make reverts as easy as a key click, and reward editors with a higher "score" for performing reverts have led to the rapid and ruthless excision or modification of content, with only an automated explanation to the original editor.

For this reason, University of Minnesota researchers are working with the Wikimedia Foundation to develop more "sociable" versions of cyborgs that will help human users become more effective members of the Wikipedia community. Over the next decade, it will be fascinating to observe what norms emerge for interaction between millions of human users and thousands of nonhuman social agents in a domain in which both have essential roles.

Aaron Halfaker *is a PhD student in the Department of Computer Science and Engineering at the University of Minnesota. Contact him at halfaker@cs.umn.edu.*

John Riedl, *Social Computing column editor, is a professor in the Department of Computer Science and Engineering at the University of Minnesota. Contact him at riedl@cs.umn.edu.*