Anon productivity and productive efficiency in English Wikipedia

Aaron Halfaker · ahalfaker@wikimedia.org
Wikimedia Research
Aaron Halfaker
Senior Research Scientist, Wikimedia Foundation

Think big. Measure what you can. Build better technologies.

About me
Hi. I'm Aaron Halfaker. I'm a scientist. See projects and publications below. I've been a Wikipedian since 2008. I mostly build tools and run studies, but I make edits where I can. In 2011, I started working with the Wikimedia Foundation as a research scientist. This is my staff account. See my volunteer account, User:EpochFail, for my non-staff work.

My work
I use the scientific method to explore the structure and function of open knowledge projects like Wikipedia. My goal is to identify problems/opportunities and to use this knowledge to improve/extend the socio-technical infrastructure that makes them work. See The Rise and Decline for an example of my research and Revision scoring as a service for an example of an experimental technology I'm actively working on.

Contact me
- E-mail: ahalfaker@wikimedia.org
- Website: http://halfaker.info
- Twitter: http://twitter.com/halfak
- On IRC, I go by "halfak".

Aaron Halfaker
Senior Research Scientist, Wikimedia Foundation

Think big. Measure what you can. Build better technologies.

Measure what you can.

About me
Hi. I'm Aaron Halfaker. I'm a scientist. See projects and publications below. I've been a Wikipedian since 2008. I mostly build tools and run studies, but I make edits where I can. In 2011, I started working with the Wikimedia Foundation as a research scientist. This is my staff account. See my volunteer account, User:EpochFail, for my non-staff work.

My work
I use the scientific method to explore the structure and function of open knowledge projects like Wikipedia. My goal is to identify problems/opportunities and to use this knowledge to improve/extend the socio-technical infrastructure that makes them work. See The Rise and Decline for an example of my research and Revision scoring as a service for an example of an experimental technology I'm actively working

Contact me
- E-mail: ahalfaker@wikimedia.org
- Website: http://halfaker.info
- Twitter: http://twitter.com/halfak
- On IRC, I go by "halfak".

Take-aways
Take-aways

1. Content persistence is a robust measurement of article-writing productivity
Take-aways

1. Content persistence is a robust measurement of article-writing productivity
2. English Wikipedia seems to be getting more efficient
Take-aways

1. Content persistence is a robust measurement of article-writing productivity
2. English Wikipedia seems to be getting more efficient
3. Anon contributions are important (15-20% of the overall productivity)
Content persistence
<table>
<thead>
<tr>
<th>Revisions</th>
<th>PWR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Apples are red.</td>
<td>6</td>
</tr>
<tr>
<td>2: Apples are blue.</td>
<td>0</td>
</tr>
<tr>
<td>3: Apples are red.</td>
<td>0</td>
</tr>
<tr>
<td>4: Apples are tasty and red.</td>
<td>1</td>
</tr>
<tr>
<td>5: Apples are tasty and blue.</td>
<td>0</td>
</tr>
<tr>
<td>Revisions</td>
<td>PWR</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>-----</td>
</tr>
<tr>
<td>1: Apples are red.</td>
<td>6</td>
</tr>
<tr>
<td>2: Apples are blue.</td>
<td>0</td>
</tr>
<tr>
<td>3: Apples are red.</td>
<td>0</td>
</tr>
<tr>
<td>4: Apples are tasty and red.</td>
<td>1</td>
</tr>
<tr>
<td>5: Apples are tasty and blue.</td>
<td>0</td>
</tr>
</tbody>
</table>

Reverted!
Difference algorithms!
Difference algorithms!
Difference algorithms!

Human intuition
Difference algorithms!

Longest common substring (e.g. unix diff)
Battle of Talienzhuang against Japanese Tanks.jpg|thumb|right|Chinese suicide bomber putting on an explosive vest made out of Model 24 hand grenades to use in an attack on Japanese tanks at the [[Battle of Talienzhuang]].

Suicide bombing was also used against the Japanese. A Chinese soldier detonated a grenade vest and killed 20 Japanese soldiers at [[Defense_of_Shiang_Warehouse#29_October|Shiang Warehouse]]. Chinese troops [[Explosive belt|strapped explosives like grenade packs or dynamite to their bodies]] and threw themselves under Japanese tanks to blow them up. In one incident at Talienzhuang, Chinese suicide bombers obliterated a column of Japanese tanks with hand-grenades.

In one incident at Talienzhuang, Chinese suicide bombers obliterated a column of Japanese tanks with hand-grenades.
Attributing authorship of content?
Attributing authorship of content?

Fabian Flöck, Maribel Acosta
*WikiWho: Precise and Efficient Attribution of Authorship of Revisioned Content*
Proceedings of the 23rd international conference on World Wide Web, ACM, April, 2014

L. de Alfaro and M. Shavlovsky.
*Attributing authorship of revisioned content.*
Proceedings of the 22nd international conference on World Wide Web, ACM, April, 2013
Segment matcher
Segment matcher

Paragraph
sentence
sentence
sentence
sentence

Paragraph
sentence
sentence
sentence

Paragraph
sentence
sentence
sentence

Paragraph
sentence
sentence
sentence
sentence
Segment matcher

sentence

sentence

sentence

sentence

sentence

sentence

sentence

sentence
Difference algorithms!

Segment Matcher == Human intuition
<table>
<thead>
<tr>
<th>Revisions</th>
<th>PWR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Apples are red.</td>
<td>6</td>
</tr>
<tr>
<td>2: Apples are blue.</td>
<td>0</td>
</tr>
<tr>
<td>3: Apples are red.</td>
<td>0</td>
</tr>
<tr>
<td>4: Apples are tasty and red.</td>
<td>1</td>
</tr>
<tr>
<td>5: Apples are tasty and blue.</td>
<td>0</td>
</tr>
</tbody>
</table>
Sensitivity analysis

How much “persisting” is enough?
Sensitivity analysis

Revisions by other people

![Graph showing revisions by other people over time](image-url)
Sensitivity analysis

Revisions by other people

Figure 1. Scatterplot of Log PWR versus mean Amazon Turk Rating

Sensitivity analysis

Sensitivity analysis

Revisions by other people

Time after save

Cutoff
“Sensitivity analysis”

Revisions by other people

Time after save

Cutoff

hazard

Revisions persisted

Time visible

step

step

Hazard of permanent removal
Sensitivity analysis

Revisions by other people

Time after save

Cutoff

Cutoff
In plain English:

Persisting word:

A *word-like* token added to an article that survives through at least 5 *revisions* by other people or 48 *hours* before being removed forever.
In plain English:

Persisting word:

A word-like token added to an article that survives through at least 5 revisions by other people or 48 hours before being removed forever.

Not complete:

Misses: Talk page activity, Template work, Image uploads, Counter-vandalism, Research & tool development, etc.
In plain English:

Persisting word:

A *word-like* token added to an article that survives through at least **5 revisions** by other people or **48 hours** before being removed forever.

Not complete:

Misses: Talk page activity, Template work, Image uploads, Counter-vandalism, Research & tool development, etc.

But good:

Recognizes: Adding good new content to articles.
Productivity Measurements!
(English Wikipedia)
Active editors

Productivity (persisting words added)
Labor hours (input)

Persisting words added (output)
2006:

155m words / 600k hours = 258 persisting words per hour

2015:

145m words / 350k hours = 414 persisting words per hour
Where’s this efficiency coming from?
Where’s this efficiency coming from?

Hypothesis: Wikipedians have better tools to help them edit articles faster.

… look for bots and automated editing tools.
No real decline here?
Definitely declining
Anons add 15-20% of productive article content.
Screenshot of AutoWikiBrowser censored due to Commons copyright policy.
reFill

Expand bare references with ease

<ref>http://example.com</ref>
→<ref>{{cite web|url=http://example.com|title=Example Domain|publisher=}}</ref>

Recent changes
- The tool is now capable of generating localized templates. Please help translate reFill into your language by joining the Transfer project. Thank you for using reFill!
- The tool can now expand New York Times references.

Fetch content from a wiki

Page name: [ ]

Options
- [ ] Use plain formatting instead of {{cite web}}
- [ ] Do not remove link rot tags
- [ ] Add blank metadata fields when the information is unavailable
- [ ] Do not add access dates
- [ ] Use the base domain name as work when this information cannot be parsed
AutoEd is a user script that helps to automatically make certain changes in articles, and it also allows for easy design, use, and customization of user scripts related to automated article cleanup. It is based on CodeFixer, Plastikspork's script, and Formatter, but allows for further customization and the easier creation of new functions. It is currently maintained primarily by Drilltho and Plastikspork.

Installation guide

It is very easy to install AutoEd and you don't need a shred of technical ability! If you can follow a recipe for making a peanut butter and jelly sandwich, you can install script. Instructions are below. If you get anxious, ask on the discussion page for help.

Efficiency up ~87%
Registered editors editing manually

Efficiency up 87%
Registered editors editing manually

Efficiency up 87%

Anons contribute 15-20%

Tool use on the rise
Registered editors editing manually

Efficiency up 87%

Anons contribute ~20%

Tool use on the rise

Mostly force multipliers and reference cleanup
How does this look for individuals?
GroupLens Research

From Wikipedia, the free encyclopedia

GroupLens Research is a human–computer interaction research lab in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities specializing in recommender systems and online communities. GroupLens also works with mobile and ubiquitous technologies, digital libraries, and local geographic information systems.

The GroupLens lab was one of the first to study automated recommender systems with the construction of the “GroupLens” recommender, a Usenet article recommendation engine, and Movielens, a popular movie recommendation site used to study recommendation engines, tagging systems, and user interfaces. The lab has also gained notability for its members' work studying open content communities such as Wikipedia and Cyclopedia, a computational "geo-wiki" currently being used in the Twin Cities to help plan the regional cycling system.\[1\]

Contents
- [History](#)
- [Contributions](#)
- [References](#)
Non-self persisting tokens

month

Tools I've developed
- ORES is a AI prediction as a service. It hosts machine learning models for predicting vandalism, article quality, etc.
- Wiki labels is a generalized, crowdsourced labeling system for Wikipedia. (see also en:WP:Labels)
- Snuggle supports wiki-mentors by allowing them to find desirable newcomers that need mentorship.
- Mr. Clean allows readers to insert cleanup templates in articles.
- Wikignome allows edits to individual sentences while viewing an article.
- NICE is an experimental interface that asks reverting editors to be nice to newbies.
- HAPPI adds highlighting to edit pane to visually present information about who last editing wiki-markup and how long ago.
- NOOB adds a visualization of editor experience to talk pages and the revision history page.
- APS Review is a review system for APS.WI to allow members to more easily submit reviews of the articles that they read.
Guillom
Guillom
Take-aways
Take-aways

1. Content persistence is a robust measurement of article-writing productivity

<table>
<thead>
<tr>
<th>Revisions</th>
<th>PWR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Apples are red.</td>
<td>6</td>
</tr>
<tr>
<td>2: Apples are blue.</td>
<td>0</td>
</tr>
<tr>
<td>3: Apples are red.</td>
<td>0</td>
</tr>
<tr>
<td>4: Apples are tasty and red.</td>
<td>1</td>
</tr>
<tr>
<td>5: Apples are tasty and blue.</td>
<td>0</td>
</tr>
</tbody>
</table>

Survival of content → Quality

Tracking authorship is hard

Useful measure, but incomplete

Tools I’ve developed
- ORES is a AI prediction as a service. It hosts machine learning models for predictions.
- Wiki labels is a generalized, crowdsourced labeling system for Wikipedia. (see...)
- Snuggle supports wiki-mentors by allowing them to find desirable newcomers.
- Mr. Clean allows readers to insert cleanup templates in articles.
- Wikignome allows edits to individual sentences while viewing an article.
- NICE is an experimental interface that asks reverting editors to be nice to new users.
- HAPPI adds highlighting to edit pane to visually present information about who is what.
- NOOB adds a visualization of editor experience to talk pages and the revision...
Take-aways

1. Content persistence is a robust measurement of article-writing productivity
2. English Wikipedia seems to be getting more efficient

+86% efficiency since 2006!
Take-aways

1. Content persistence is a robust measurement of article-writing productivity
2. English Wikipedia seems to be getting more efficient
3. Anon contributions are important
   15-20% of the overall productivity!
Next steps
More wikis

- Processing English Wikipedia is most difficult -- smaller wikis are easier

- Target “emerging communities” -- see https://meta.wikimedia.org/wiki/Community_Resources
Productivity

1: Apples are red.
2: Apples are blue.
3: Apples are red.
4: Apples are tasty and red.
5: Apples are tasty and blue.

Importance

Page views
Link graph:
- Indegree
- Page rank

“Value-added”
Lorem ipsum dolor sit amet, consectetur adipiscing elit.

In molestie justo vel enim
In molestie justo vel enim vestibulum consectetur.

User: ValueAdder

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edits</td>
<td>1,230</td>
</tr>
<tr>
<td>Edit sessions</td>
<td>~27</td>
</tr>
<tr>
<td>Time spent editing</td>
<td>~225 hours</td>
</tr>
<tr>
<td>Content added</td>
<td>3.5 MB</td>
</tr>
<tr>
<td>Word persistence</td>
<td>93.7%</td>
</tr>
</tbody>
</table>

Top contributions

<table>
<thead>
<tr>
<th>Title</th>
<th>PWI</th>
<th>Contrib</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lorem ipsum</td>
<td>765</td>
<td>2.5k</td>
</tr>
<tr>
<td>Dolor sit</td>
<td>353</td>
<td>1.2k</td>
</tr>
<tr>
<td>Amet get auctor</td>
<td>123</td>
<td>987</td>
</tr>
<tr>
<td>Molestie nulla</td>
<td>85</td>
<td>778</td>
</tr>
<tr>
<td>Ligula</td>
<td>75</td>
<td>342</td>
</tr>
<tr>
<td>Luctus, Velit (po...)</td>
<td>22</td>
<td>100</td>
</tr>
<tr>
<td>Aenean</td>
<td>10</td>
<td>77</td>
</tr>
</tbody>
</table>

Contact us


Content persistence
In molestie justo vel enim vestibulum consectetur.

Sharing and privacy
Aenean eleifend eu turpis et pretium. In at imperdiet orci. Proin id vulputate ligula, non vehicula tellus. Proin id vulputate ligula, non vehicula tellus.
Open Access Datasets

Quarry

EpochFall's edits in Enwiki since 20140101

Query status: complete
Results: 1 row(s)

https://commons.wikimedia.org/wiki/File:Open_Access_logo_PLoS_transparent.svg (CC0)

https://commons.wikimedia.org/wiki/File:Quarry-logo.svg (CC-BY-SA 4.0)
Thanks!

Aaron Halfaker
ahalfaker@wikimedia.org

[[User:EpochFail]] & [[User:Halfak (WMF)]]

“halfak” on IRC

@halfak on Twitter

https://github.com/halfak