Halfak’s Wiki Wiki Research
Libraries
v 0.0.1
Outline

1. Openness: Data, algorithms & results
2. Current Libraries
   a. pip install mediawiki-utilities
   b. pip install wikiclass
   c. pip install deltas
   d. pip install mwoauth
Part 1

What is open research?
Three major artifacts
Three major artifacts

1. The manuscript: “Open Access”
The Rise and Decline of an Open Collaboration System: How Wikipedia’s Reaction to Popularity Is Causing Its Decline

Aaron Halfaker  aaron.halfaker@gmail.com
R. Stuart Geiger
Jonathan T. Morgan
John Riedl

Abstract

This Article
Published online before print
American Behavioral Scientist
December 28, 2012
002764212469365

» Abstract Free
Full Text (PDF)

All Versions of this Article:
Version of Record · Apr 8, 2013
OnlineFirst Version of Record · Dec 28, 2012
What's this?

Current Issue
August 2014, 58 (9)

Alert me to new issues of American Behavioral Scientist
The Rise and Decline of an Open Collaboration System: How Wikipedia’s Reaction to Popularity Is Causing Its Decline

Aaron Halfaker  aaron.halfaker@gmail.com
R. Stuart Geiger
Jonathan T. Morgan
John Riedl

Abstract

Full Text (PDF)
You may purchase this article for US$30.00.
My hack

Open access summary

(I am not a lawyer)
My current process

Research: Wikipedia article creation

The process of creating articles is becoming increasingly difficult for new users due to increasingly restrictive criteria\(^1\) and the speed at which their articles are tagged and deleted\(^2\). This trend is concerning because new users tend to leave the wiki when their work is deleted.

The English Wikipedia Articles for Creation WikiProject has recently adjusted in order to encourage new editors to create draft articles outside of the usual article space. However, it's unclear whether such initiatives are successful in improving the success rate of articles created by new editors or improving their retention. In this study, we'll discuss our analysis of newcomer created articles in the most active Wikipedia projects and answer questions about how different workflows affect the success rate of articles.

Related work \[edit\]

Research has established that the number of active editors in the English Wikipedia has entered a decline and that this decline is the result of decreased retention of new users\(^3\). Subsequent research by Halfaker et al. has shown evidence that this decline is not due to the quality of newcomers, but rather the increasing complexity newcomers must manage in order to successfully contribute and the negative reactions they receive\(^4\). One of the key factors in Halfaker et al.'s model predicting the retention of new editors was whether they created articles that were quickly deleted. Related work by User:Mr.Z-man confirmed that new editors who created articles that were deleted are less likely to continue to contribute\(^5\). Research performed in parallel found that the rate at which newly created articles are deleted has risen sharply in recent years\(^6\) and the speed at which new articles are tagged and deleted has increased dramatically\(^2\).
Three major artifacts

1. The manuscript: “Access”
2. The data: “Data”
Data dumps

Summary

WMF publishes data dumps of Wikipedia and all WMF projects on a regular basis. English Wikipedia is dumped once a month, while smaller projects are often dumped twice a month.

Description

Text and metadata of current or all revisions of all pages as XML files
Most database tables as sql files
- Page-to-page link lists (pagelinks, categorylinks, imagelinks, templateslinks tables)
- Lists of pages with links outside of the project (externallinks, iwllinks, langlinks tables)
- Media metadata (image, oldimage tables)
- Info about each page (page, page_props, page_restrictions tables)
- Titles of all pages in the main namespace, i.e. all articles (*.all-titles-in-ns0.gz)
- List of all pages that are redirects and their targets (redirect table)
A collection of datasets about Wikipedia and other projects run by the Wikimedia Foundation. The collection is open to contributions by researchers not affiliated with

15 datasets found

Teahouse corpus

The Teahouse corpus is a set of questions asked at the Wikipedia Teahouse, a peer support forum for new Wikipedia editors. This corpus contains data from its first two years of...
See also

- Teahouse project documentation: project planning docs and reports from the Teahouse pilot
- Wikimedia data portal: public data resources on Wikipedia and other Wikimedia projects
- Mediawiki database schema: description of standard data tables and fields in MediaWiki sites
- Mediawiki API documentation: data available through the MediaWiki API (depending on site configuration)
- Wikitext markup information: information about the markup conventions used in the text of the Teahouse corpus

Data and Resources

- **Teahouse questions - 2/23/2014**
  Metadata for 5,003 questions. See README: Teahouse questions for field...

- **Teahouse question text - 2/23/2014**
  The raw text of 4,998 questions. See README: Teahouse question text for...

- **README: Teahouse question text**
  Field types and value definitions for teahouse-question-text datafile

- **README: Teahouse questions**
  Data field types and values
Three major artifacts

1. The manuscript: “Access”
2. The data: “Data”
3. The code: “Source”
Why open source?

Red Hat® believes open source simply creates better software. Everyone collaborates. The best technology wins. Not just within one company, but for everyone, anyone, around the world. [http://www.redhat.com/about/whoisredhat/opensource.html](http://www.redhat.com/about/whoisredhat/opensource.html)

... but for science?
Why open source?

HALFAK believes open source simply creates better SCIENCE. Everyone collaborates. The best SCIENCE HAPPENS. Not just within one company, but for everyone, anyone, around the world. --Halfak

… but for science?
Part 2
Libraries!
Libraries

- **MediaWiki Utilities** -- General data processing
  - Repo: https://github.com/halfak/Mediawiki-Utilities
  - Docs: http://pythonhosted.org/mediawiki-utilities

- **Wiki-Class** -- Article quality classification
  - Repo: https://github.com/halfak/Wiki-Class
  - Docs: https://pythonhosted.org/wikiclass

- **MediaWiki OAuth** -- OAuth handshaker
  - Repo: https://github.com/halfak/MediaWiki-OAuth
  - Docs: http://pythonhosted.org/mwoauth

- **Deltas** -- Robust difference detection
  - Repo: https://github.com/halfak/Deltas
  - Docs: http://pythonhosted.org/deltas
OAuth!

Allows Wiki-tool users to log into Wikipedia with their accounts.
OAuth Handshake

1. Initialize
   - User
     - Your tool (client consumer)
     - MediaWiki (provider)

2. Authorize
   - User
   - Your tool (client consumer)
   - MediaWiki (provider)

3. Complete
   - Your tool (client consumer)
   - MediaWiki (provider)
OAuth Handshake

1. Initialize

Your tool
consumer

User
resource
owner

MediaWiki
provider

Request token

2. Authorize

Your tool
consumer

User
resource
owner

MediaWiki
provider

Verifier

3. Complete

Your tool
consumer

User
resource
owner

MediaWiki
provider

Access token
OAuth Handshake

1. Initialize
2. Authorize
3. Complete

```
index.php?
title=Special%3AOAuth%2Fauthorize&
oauth_consumer_key=<request key>&
oauth_token=<request secret>
```
OAuth Handshake

1. Initialize

2. Authorize

index.php?
title=Special%3AOAuth%2Fauthorize&
oauth_consumer_key=<request key>&
oauth_token=<request secret>

Request token

Verifier

JSON Web Tokens

Table of Contents

1. Introduction
   1.1. Notational Conventions
2. Terminology
3. JSON Web Token (JWT) Overview
   3.1. Example JWT
4. JWT Claims
   4.1. Registered Claim Names
      4.1.1. "iss" (Issuer) Claim
      4.1.2. "sub" (Subject) Claim
      4.1.3. "aud" (Audience) Claim
      4.1.4. "exp" (Expiration Time) Claim
      4.1.5. "nbf" (Not Before) Claim
      4.1.6. "iat" (Issued At) Claim
      4.1.7. "jti" (JWT ID) Claim
   4.2. Public Claim Names
   4.3. Private Claim Names
   4.4. JOSE Header
      4.4.1. "typ" (Type) Header Parameter
      4.4.2. "cty" (Content Type) Header Parameter
   5. Regrading Claims as Header Parameters
   5.3. Plaintext JWTs
   5.4. Example Plaintext JWT
7. Rules for Creating and Validating a JWT
7.1. String Comparison Rules
8. Implementation Requirements
9. URI for Declaring that Content is a JWT
10. IANA Considerations
16.1. JSON Web Token Claims Registry
16.1.1. Registration Template
16.1.2. Initial Registry Contents
19.2. Sub-Namespace Registration of urn:ietf:params:oauth:token-type:jwt
19.3. Media Type Registration
19.3.1. Registry Contents
19.4. Header Parameter Names Registration
19.4.1. Registry Contents
11. Security Considerations
11.1. Trust Decisions
11.2. Signing and Encryption Order
13. Privacy Considerations
13.1. References
13.2. Informative References
Appendix A. JWT Examples
A.1. Example Encrypted JWT
A.2. Example Nested JWT
Appendix B. Relationship of JWEs to SAML Assertions
Appendix C. Relationship of JWEs to Simple Web Tokens (SWTs)
Appendix D. Acknowledgements
Appendix E. Document History
3. Authors' Addresses
OAuth Handshake

1. Initialize
2. Authorize
3. Complete

Request token
Verifier
Access token

index.php?
title=Special%3AOAuth%2Fauthorize&
oauth_consumer_key=<request key>&
oauth_token=<request secret>

Check the **issuer**
Confirm that token.iss matches the domain of MediaWiki that you made the request to (e.g. "mediawiki.org")

Check the **audience**
Confirm that token.aud matches your consumer key

Check the **issued at time**
Confirm that token.iat (unix timestamp in seconds) is before the current time

Check the **expiration time**
Confirm that token.exp (unix timestamp in seconds) is after the current time

Check the **number used only once**
Confirm that token.nonce matches the nonce your application sent with the original request

JSON Web Tokens

Table of Contents

1. Introduction
2. Notational Conventions
3. Terminology
4. JSON Web Token (JWT) Overview
5. Example JWT
6. JWT Claims
7. Registered Claim Names
   7.1. "iss" (Issuer) Claim
   7.2. "sub" (Subject) Claim
   7.3. "aud" (Audience) Claim
   7.4. "exp" (Expiration Time) Claim
   7.5. "nbf" (Not Before) Claim
   7.6. "iat" (Issued At) Claim
   7.7. "jti" (JWT ID) Claim
8. Public Claim Names
9. Private Claim Names
10. JOSE Header
11. "typ" (Type) Header Parameter
12. "cty" (Content Type) Header Parameter
13. "claim" (Claim Set) Header Parameter
14. Jwe Claims as Header Parameters
15. Plaintext Jwts
16. Example Plaintext JWT
17. Rules for Creating and Validating a JWT
18. String Comparison Rules
19. Implementation Requirements
20. URI for Declaring that Content is a JWT
21. IANA Considerations
22. JSON Web Token Claims Registry
23. Initial Registry Contents
24. Sub-Namespace Registration of urn:oasis:names:tc:token-type:jwt
25. Registry Contents
26. Security Considerations
27. Trust Decisions
28. Security and Encryption Order
29. Privacy Considerations
30. References
   30.1. Normative References
   30.2. Informative References
Appendix A. JWT Examples
A.1. Example Encrypted JWT
A.2. Example Nested JWT
Appendix B. Relationship of Jwts to SAML Assertions
Appendix C. Relationship of Jwts to Simple Web Tokens (SWTs)
Appendix D. Acknowledgements
Appendix E. Document History
3. Authors' Addresses
[back to the editor]
Difference algorithms!
Difference algorithms!

Human intuition
Difference algorithms!

Longest common substring
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
Difference algorithms!

Segment Matcher == Human intuition
Research software as libraries

- Easy to re-use
  - `pip install mediawiki-utils`
Research software as libraries

● Easy to re-use
  ○ pip install mediawiki-utilities

● More collaboration == more science
  ○ http://github.com/halfak/mediawiki-utilities
  ○ plz submit bugs and pull requests!
Research software as libraries

● Easy to re-use
  ○ pip install mediawiki-utilities

● More collaboration == more science
  ○ http://github.com/halfak/mediawiki-utilities
  ○ plz submit bugs and pull requests!

● Broad benefit & consistency
  ○ Complex problems
  ○ Bugs
Thanks!

http://github.com/halfak/Mediawiki-Utilities
http://github.com/halfak/Wiki-class
http://github.com/halfak/MediaWiki-OAuth
http://github.com/halfak/Deltas

Send me pull requests!

Props to:
- Yuvi Panda
- Filippo Valsorda
- Max Klien
- Morten Warncke-Wang
- Oliver Keyes

Aaron Halfaker
ahalfaker@wikimedia.org
“halfak” on IRC
@halfak on Twitter