A Breakdown of Quality Flaws in Wikipedia

Maik Anderka Benno Stein

Bauhaus-Universität Weimar
www.webis.de

WebQuality 2012
Lyon, France
April 16, 2012

Wikipedia Facts

- Free online encyclopedia
- Launched in January 2001
- 285 languages
- 81,906,954 pages
- 21,693,832 encyclopedic articles
- 1,992,206 images
- 33,940,434 registered users
- 4,574 admins
- `wikipedia.org` is the sixth most-visited website

[http://www.alexa.com/siteinfo/wikipedia.org]

[http://meta.wikimedia.org/wiki/List_of_Wikipedias]

What about Information Quality?

What about Information Quality?

- Everyone can edit Wikipedia, even anonymously
- Heterogeneous community of Wikipedia authors
- Edits are not reviewed before publication

➜ **Inequality in content quality**

What about Information Quality?

- Everyone can edit Wikipedia, even anonymously
- Heterogeneous community of Wikipedia authors
- Edits are not reviewed before publication

➜ **Inequality in content quality**

Problems:

- Improving low-quality content
- Maintaining high-quality content

Automatic Quality Assessment

- Up to now: classification into abstract quality schemes
- For instance “Is an article featured or not?”

  [Likpa and Stein, WWW 2010], [Dalip et al., JCDL 2009], [Blumenstock, WWW 2008], [Hu et al., CIKM 2007], [Wilkinson and Huberman, WikiSym 2007]
- Performs nearly perfect, but
  - No rationale why an article violates Wikipedia’s featured article criteria
  - No practical support for Wikipedia’s quality assurance process

Automatic Quality Assessment

- Up to now: classification into abstract quality schemes
- For instance “Is an article featured or not?”

  [Likpa and Stein, WWW 2010], [Dalip et al., JCDL 2009], [Blumenstock, WWW 2008], [Hu et al., CIKM 2007], [Wilkinson and Huberman, WikiSym 2007]

- Performs nearly perfect, but
  - No rationale why an article violates Wikipedia’s featured article criteria
  - No practical support for Wikipedia’s quality assurance process

- Less than 0.1% of the English Wikipedia articles are featured

➜ **What is wrong with the remaining 99.9%?**

We investigate *specific quality flaws* in Wikipedia

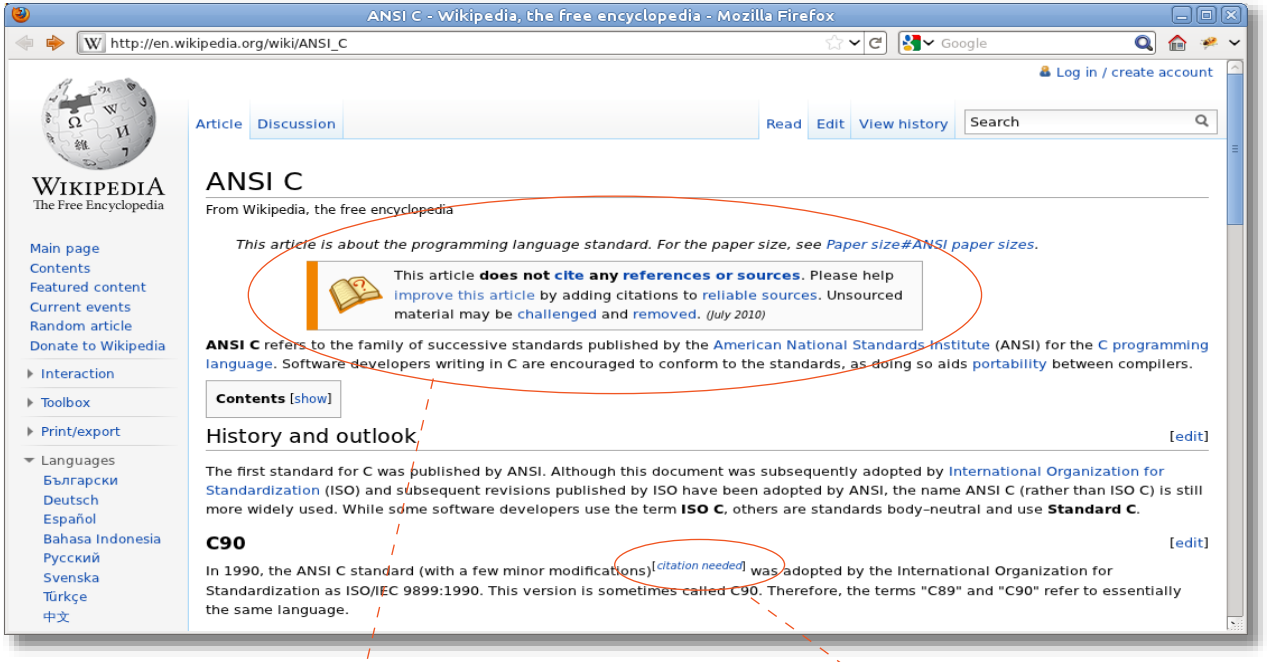
Outline

- Motivation
- Compiling Existing Quality Flaws
- Quality Flaw Organization
- Measuring the Extend of Flawed Content
- Current Work: Quality Flaw Prediction
- Summary

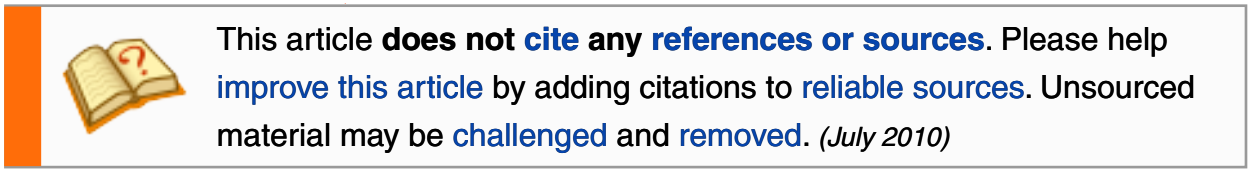
What quality flaws actually occur in Wikipedia?

Compiling Existing Quality Flaws

Cleanup Tags



[*citation needed*]



Compiling Existing Quality Flaws

Cleanup Tags

- Idea: each cleanup tag defines a certain quality flaw

Problem:

- Cleanup tags are realized based on templates
- The English Wikipedia contains more than 320,000 templates
  - ➜ Automated cleanup tag mining

Cleanup Tag Mining

Data base

- English Wikipedia snapshot from January 15, 2011

Preprocessing

- Importing the SQL dumps (>40GB) into a local Wikipedia database

Mining approach

1. Extract candidate cleanup tags from two meta sources:
   - Administration category *Category:Cleanup_templates*
   - Meta page *Wikipedia:Template_messages/Cleanup*
2. Refinement:
   - Resolve redirects
   - Discard meta-templates, documentation-, and test pages

Result

- 388 cleanup tags

Outline

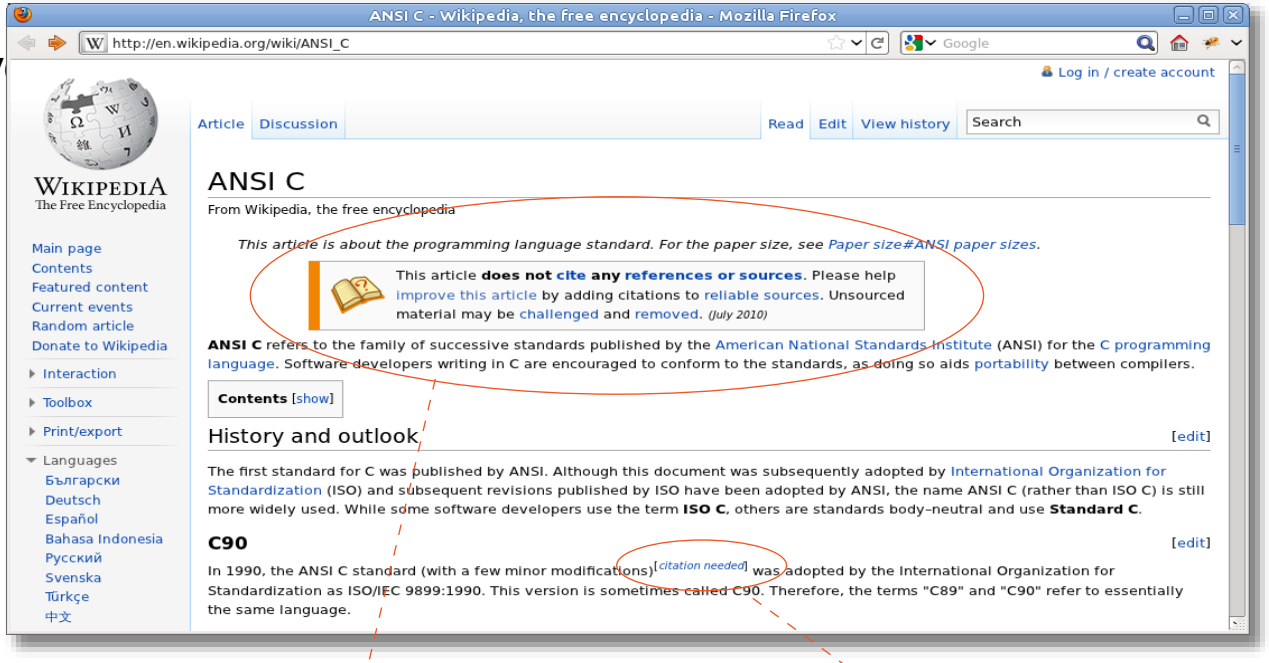
- Motivation
- Compiling Existing Quality Flaws
- Quality Flaw Organization
- Measuring the Extend of Flawed Content
- Current Work: Quality Flaw Prediction
- Summary

Quality Flaw Organization – Type

- Several flaws relate to the same type

Quality Flaw Organization – Type

- Sev



[*citation needed*]

This article **does not cite any references or sources**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(July 2010)*

Quality Flaw Organization – Type

- Several flaws relate to the same type
- We organize the flaws along 12 general flaw types

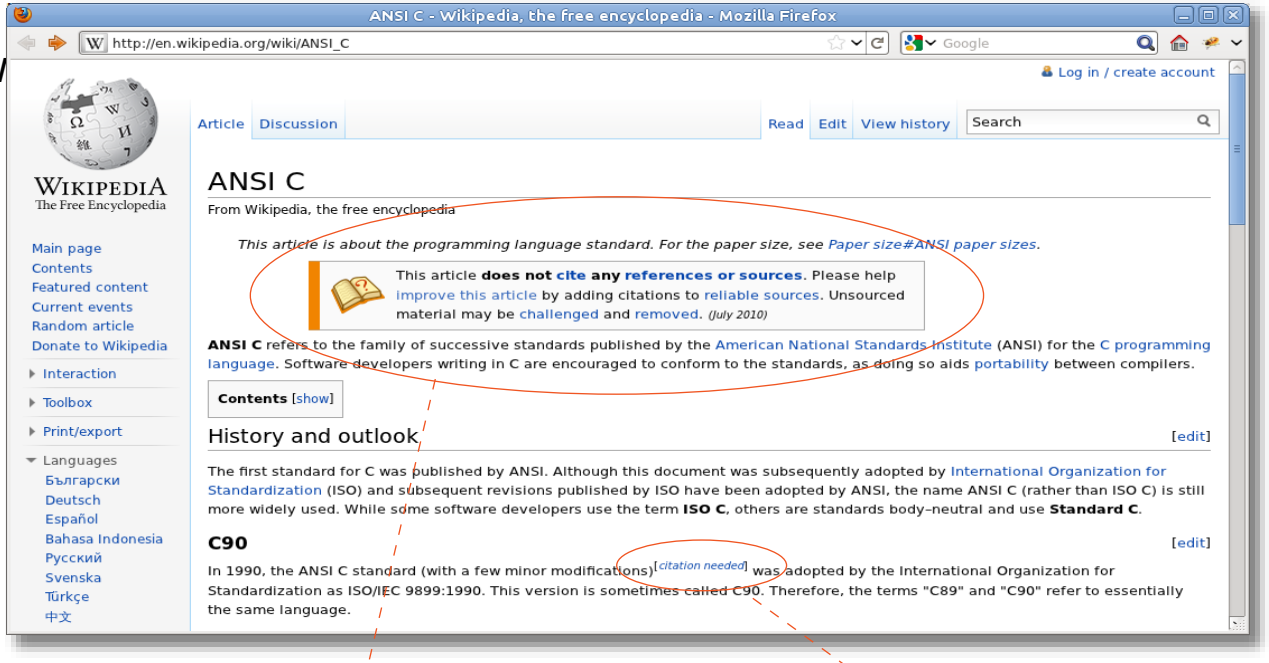
Flaw type	Cleanup tags
Verifiability	66
Style of writing	66
Cleanup of specific subjects	60
Miscellaneous	50
Neutrality	35
Unwanted content	33
Wiki tech	21
General cleanup	18
Structure	16
Expand	11
Merge	6
Time-sensitive	6
$\sum$	388

Quality Flaw Organization – Scope

- Flaws differ by their scope

Quality Flaw Organization – Scope

- Flaw



[*citation needed*]

This article **does not cite any references or sources**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(July 2010)*

Quality Flaw Organization – Scope

- Flaws differ by their scope
- We distinguish two scopes

Scope	Cleanup tags
Article flaws (tag boxes)	307
Inline flaws (inline tags)	81
$\sum$	388

Outline

- Motivation
- Compiling Existing Quality Flaws
- Quality Flaw Organization
- Measuring the Extend of Flawed Content
- Current Work: Quality Flaw Prediction
- Summary

Where do quality flaws occur?

Measuring the Extend of Flawed Content

Distribution over Namespaces

Namespace	Pages	Tagged pages	Ratio in %
Main (articles)	3 557 468	979 299	27.53
File	875 833	32 184	3.67
Wikipedia	607 854	2 321	0.38
Template	329 282	544	0.17
...	...	...	...
Talk	3 901 638	121 717	3.12
User	1 118 456	8 586	0.77
User talk	6 377 234	3 366	0.05
Wikipedia talk	84 996	733	0.86
...	...	...	...
	$\sum$ 22 981 145	$\sum$ 1 149 428	Ø 5.00

Measuring the Extend of Flawed Content

Distribution over Namespaces

Namespace	Pages	Tagged pages	Ratio in %
Main (articles)	3 557 468	979 299	27.53
File	875 833	32 184	3.67
Wikipedia	607 854	2 321	0.38
Template	329 282	544	0.17
...	...	...	...
Talk	3 901 638	121 717	3.12
User	1 118 456	8 586	0.77
User talk	6 377 234	3 366	0.05
Wikipedia talk	84 996	733	0.86
...	...	...	...
	$\sum$ 22 981 145	$\sum$ 1 149 428	Ø 5.00

- ➜ The majority of flaws is tagged in the encyclopedic content
- ➜ 1 in 4 articles is tagged with a quality flaw

What topics are likely to be tagged?

Measuring the Extend of Flawed Content

Distribution over Main Topics

Topic	Articles	Tagged articles	Ratio, %
Computers	46 897	22 748	48.51
Belief	23 084	10 695	46.33
...	...	...	...
Chronology	905 090	237 122	26.20
Mathematics	23 499	5 974	25.42
Geography	726 938	144 157	19.83

Measuring the Extend of Flawed Content

Distribution over Main Topics

Topic	Articles	Tagged articles	Ratio, %
Computers	46 897	22 748	48.51
Belief	23 084	10 695	46.33
...	...	...	...
Chronology	905 090	237 122	26.20
Mathematics	23 499	5 974	25.42
Geography	726 938	144 157	19.83

➜ Controversial topics are more likely to be tagged

What are the most frequent quality flaw types?

Measuring the Extend of Flawed Content

Flaw Frequency by Type

Flaw type	Cleanup tags	Tagged articles	Ratio in %
Verifiability	66	692 241	19.46%
Wiki tech	21	194 649	5.47%
General cleanup	18	71 401	2.01%
Expand	11	64 450	1.81%
Unwanted content	33	51 130	1.44%
Style of writing	66	42 972	1.21%
Neutrality	35	18 023	0.51%
Merge	6	15 251	0.43%
Cleanup of specific subjects	60	7 474	0.21%
Structure	16	7 280	0.20%
Time-sensitive	6	6 185	0.17%
Miscellaneous	50	2 208	0.06%
	$\sum$ 388	$\sum$ 979 187	Ø 27.52%

Measuring the Extend of Flawed Content

Flaw Frequency by Type

Flaw type	Cleanup tags	Tagged articles	Ratio in %
Verifiability	66	692 241	19.46%
Wiki tech	21	194 649	5.47%
General cleanup	18	71 401	2.01%
Expand	11	64 450	1.81%
Unwanted content	33	51 130	1.44%
Style of writing	66	42 972	1.21%
Neutrality	35	18 023	0.51%
Merge	6	15 251	0.43%
Cleanup of specific subjects	60	7 474	0.21%
Structure	16	7 280	0.20%
Time-sensitive	6	6 185	0.17%
Miscellaneous	50	2 208	0.06%
	$\sum$ 388	$\sum$ 979 187	Ø 27.52%

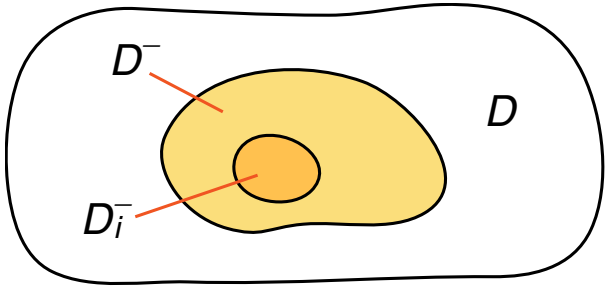
➜ 70% of the tagged flaws concern verifiability of information

But: The number of tagged articles does not show the whole picture

Measuring the Extend of Flawed Content

Actual Flaw Frequency

- It is more than likely that many flaws are not yet identified



D = English Wikipedia articles

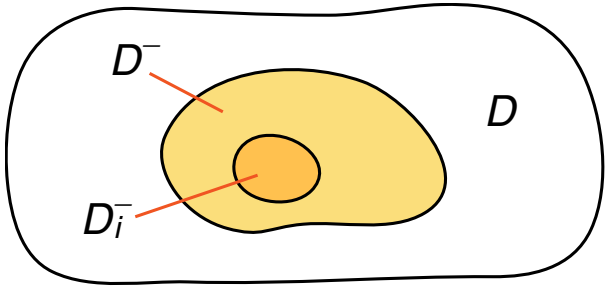
D^{-} = Articles tagged with at least one flaw

D_i^{-} = Articles tagged with flaw f_i

Measuring the Extend of Flawed Content

Actual Flaw Frequency

- It is more than likely that many flaws are not yet identified



D = English Wikipedia articles

D^- = Articles tagged with at least one flaw

D_i^- = Articles tagged with flaw f_i

➜ The number of tagged articles is just a lower bound

➜ The actual number of flaws is even higher

Outline

- Motivation
- Compiling Existing Quality Flaws
- Quality Flaw Organization
- Measuring the Extend of Flawed Content
- Current Work: Quality Flaw Prediction
- Summary

Current Work: Quality Flaw Prediction

Problem Statement

Given a set of Wikipedia articles that are tagged with a particular quality flaw, decide whether an untagged article suffers from this flaw.

Current Work: Quality Flaw Prediction

Problem Statement

Given a set of Wikipedia articles that are tagged with a particular quality flaw, decide whether an untagged article suffers from this flaw.

Approach [Anderka et al., CIKM 2011] [Anderka et al., WWW 2011]

- Train a one-class classifier for each flaw
- Document model: more than 100 Wikipedia article features
- 4 from 10 important flaws can be detected with a precision close to 1

Current Work: Quality Flaw Prediction

Problem Statement

Given a set of Wikipedia articles that are tagged with a particular quality flaw, decide whether an untagged article suffers from this flaw.

Approach [Anderka et al., CIKM 2011] [Anderka et al., WWW 2011]

- Train a one-class classifier for each flaw
- Document model: more than 100 Wikipedia article features
- 4 from 10 important flaws can be detected with a precision close to 1

PAN @ CLEF'12

- Competition on Quality Flaw Prediction in Wikipedia
- `http://pan.webis.de`

Outline

- Motivation
- Compiling Existing Quality Flaws
- Quality Flaw Organization
- Measuring the Extend of Flawed Content
- Current Work: Quality Flaw Prediction
- Summary

Summary

What we have done

- Automated cleanup tag mining approach
- Organization of quality flaws into flaw type and scope
- Breakdown of the extend of flawed content in terms of tagged articles per namespaces, main topics, and flaw types

Summary

Take away messages & key points

- 1 in 4 English Wikipedia articles is tagged to contain a quality flaw
- 70% of the tagged flaws concern article verifiability
- The actual number of flaws is even higher
- The majority of flaws is tagged in the encyclopedic content
- Controversial topics are more likely to be tagged

Thank you!

maik.anderka@uni-weimar.de