

Automatic Detection of Information Quality Flaws in Wikipedia Articles

Maik Anderka
Bauhaus-Universität Weimar
www.webis.de

WIQE Workshop, Valencia, 14. Sep. 2010

Outline

- Background and Previous Work
- Investigating IQ Flaws of Wikipedia Articles
- Article Quality Model
- IQ Flaw Corpus
- Current Work: IQ Flaw Classification
- Summary

What is Information Quality?

In General

Information Quality (IQ) is:

- subjective
- dependent on context
- a multidimensional concept

In Wikipedia

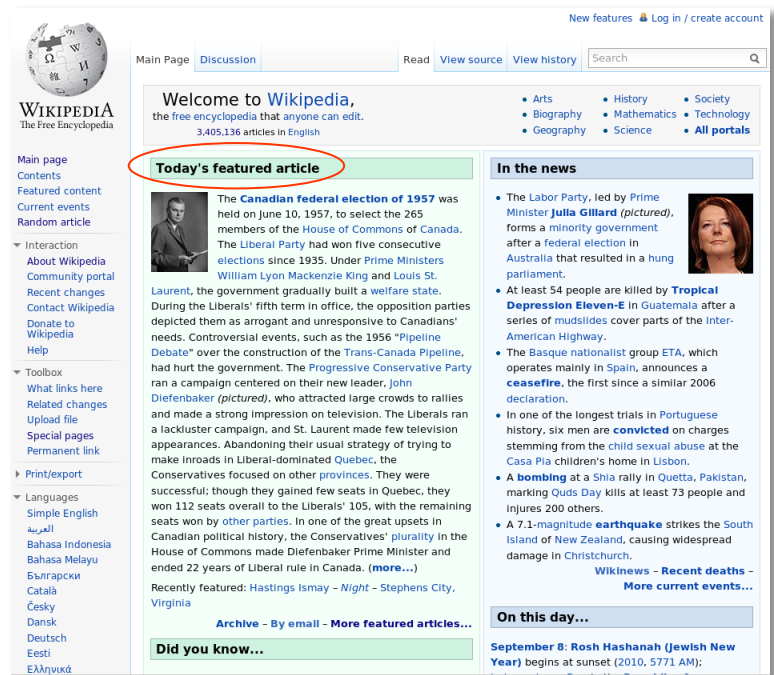
- The context is well-specified by the encyclopedic genre.
- The IQ of an article is defined by the featured article criteria.

IQ Assurance in Wikipedia

... means to guarantee that the articles fulfill a set of general IQ assessment criteria, called *featured article criteria*.

Featured articles

- ❑ The best articles in Wikipedia.
- ❑ Fulfill the featured article criteria.
- ❑ Community-driven nomination and review process.
- ❑ < 0.1 % of the English Wikipedia articles are featured.



The screenshot shows the English Wikipedia homepage. At the top, there is a search bar and navigation links. The main content area features a 'Welcome to Wikipedia' message and a 'Today's featured article' section, which is highlighted with a red circle. The featured article is about the 'Canadian federal election of 1957'. To the right, there is a 'In the news' section with several news items. At the bottom, there is a 'Did you know...' section.

WIKIPEDIA
The Free Encyclopedia

3,405,136 articles in English

Today's featured article

The **Canadian federal election of 1957** was held on June 10, 1957, to select the 265 members of the House of Commons of Canada. The Liberal Party had won five consecutive elections since 1935. Under Prime Ministers William Lyon Mackenzie King and Louis St. Laurent, the government gradually built a welfare state. During the Liberals' fifth term in office, the opposition parties depicted them as arrogant and unresponsive to Canadians' needs. Controversial events, such as the 1956 "Pipeline Debate" over the construction of the Trans-Canada Pipeline, had hurt the government. The Progressive Conservative Party ran a campaign centered on their new leader, John Diefenbaker (pictured), who attracted large crowds to rallies and made a strong impression on television. The Liberals ran a lackluster campaign, and St. Laurent made few television appearances. Abandoning their usual strategy of trying to make inroads in Liberal-dominated Quebec, the Conservatives focused on other provinces. They were successful; though they gained few seats in Quebec, they won 112 seats overall to the Liberals' 105, with the remaining seats won by other parties. In one of the great upsets in Canadian political history, the Conservatives' plurality in the House of Commons made Diefenbaker Prime Minister and ended 22 years of Liberal rule in Canada. (more...)

Recently featured: Hastings Ismay - *Night* - Stephens City, Virginia

Archive - By email - More featured articles...

Did you know...

In the news

- The Labor Party, led by Prime Minister **Julia Gillard** (pictured), forms a minority government after a federal election in Australia that resulted in a hung parliament.
- At least 54 people are killed by **Tropical Depression Eleven-E** in Guatemala after a series of mudslides cover parts of the Inter-American Highway.
- The Basque nationalist group ETA, which operates mainly in Spain, announces a **ceasefire**, the first since a similar 2006 declaration.
- In one of the longest trials in Portuguese history, six men are **convicted** on charges stemming from the child sexual abuse at the Casa Pia children's home in Lisbon.
- A **bombing** at a Shia rally in Quetta, Pakistan, marking Quds Day kills at least 73 people and injures 200 others.
- A 7.1-magnitude **earthquake** strikes the South Island of New Zealand, causing widespread damage in Christchurch.

Wikinews - Recent deaths - More current events...

On this day...

September 8: Rosh Hashanah (Jewish New Year) begins at sunset (2010, 5:71 AM);

Previous Work

Automatic IQ assessment in Wikipedia

- ❑ The Focus is almost exclusively on the classification task:
“Is an article featured or not?”
- ❑ Approaches mainly differ in
 - the machine learning algorithm,
 - the set of features, and
 - the test- and training set.
- ❑ The best approaches perform nearly perfect.
- ❑ **But:** There is little support for Wikipedia’s IQ assurance process.
 - Featured articles are not found, they are *made* by the community!

Main Idea

Automatic detection of concrete IQ flaws in Wikipedia articles

- The question is: “*What* makes a Wikipedia article a low-quality article?”

- Benefits:
 - Tells users what needs to be done to improve the IQ of an article.
 - Helps to identify flawed information.
 - Can be used to automate parts of the tagging work in Wikipedia.
 - Enables intelligent task routing.

Problem Definition

The automatic detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.

Problem Definition

The automatic detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.

- F is the set of IQ flaws occurring in Wikipedia articles.

Problem Definition

The automatic detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.

- F is the set of IQ flaws occurring in Wikipedia articles.
- D is the set of low-quality Wikipedia articles, where each $d \in D$ has at least one IQ flaw $f \in F$.

Problem Definition

The automatic detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.

- F is the set of IQ flaws occurring in Wikipedia articles.
- D is the set of low-quality Wikipedia articles, where each $d \in D$ has at least one IQ flaw $f \in F$.
- $D_c \subset D$ is a corpus containing pre-classified articles.

Problem Definition

The automatic detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.

- F is the set of IQ flaws occurring in Wikipedia articles.
- D is the set of low-quality Wikipedia articles, where each $d \in D$ has at least one IQ flaw $f \in F$.
- $D_c \subset D$ is a corpus containing pre-classified articles.
- $\alpha : D \rightarrow \mathbf{D}$ is an article quality model.

Problem Definition

The automatic detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.

- F is the set of IQ flaws occurring in Wikipedia articles.
- D is the set of low-quality Wikipedia articles, where each $d \in D$ has at least one IQ flaw $f \in F$.
- $D_c \subset D$ is a corpus containing pre-classified articles.
- $\alpha : D \rightarrow \mathbf{D}$ is an article quality model.
- $c : \mathbf{D} \rightarrow F$ is a multiclass classifier.

Problem Definition

The automatic detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.

- F is the set of IQ flaws occurring in Wikipedia articles.
- D is the set of low-quality Wikipedia articles, where each $d \in D$ has at least one IQ flaw $f \in F$.
- $D_c \subset D$ is a corpus containing pre-classified articles.
- $\alpha : D \rightarrow \mathbf{D}$ is an article quality model.
- $c : \mathbf{D} \rightarrow F$ is a multiclass classifier.
- $c : \mathbf{D} \rightarrow \mathcal{P}(F)$ is a multiclass multilabel classifier.

Problem Definition

The automatic detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.

- F is the set of IQ flaws occurring in Wikipedia articles. → ?
- D is the set of low-quality Wikipedia articles, → ?
where each $d \in D$ has at least one IQ flaw $f \in F$.
- $D_c \subset D$ is a corpus containing pre-classified articles. → ?
- $\alpha : D \rightarrow \mathbf{D}$ is an article quality model. → ?
- $c : \mathbf{D} \rightarrow F$ is a multiclass classifier. → ?

Problem Definition

The automatic detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.

- F is the set of IQ flaws occurring in Wikipedia articles. → ?
- D is the set of low-quality Wikipedia articles, where each $d \in D$ has at least one IQ flaw $f \in F$. → Previous work
- $D_c \subset D$ is a corpus containing pre-classified articles. → ?
- $\alpha : D \rightarrow \mathbf{D}$ is an article quality model. → ?
- $c : \mathbf{D} \rightarrow F$ is a multiclass classifier. → ?

Problem Definition

The automatic detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.

- F is the set of IQ flaws occurring in Wikipedia articles. → ?
- D is the set of low-quality Wikipedia articles, where each $d \in D$ has at least one IQ flaw $f \in F$. → ✓
- $D_c \subset D$ is a corpus containing pre-classified articles. → ?
- $\alpha : D \rightarrow \mathbf{D}$ is an article quality model. → ?
- $c : \mathbf{D} \rightarrow F$ is a multiclass classifier. → ?

Investigating IQ Flaws of Wikipedia Articles

Main idea

Utilize Wikipedia cleanup templates to estimate the set F of IQ flaws occurring in Wikipedia articles.

Investigating IQ Flaws of Wikipedia Articles

Wikipedia templates


Area	
- City	891.82 km ² (344.3 sq mi)
Elevation	34 - 115 m (-343 ft)
Population (2009-09-30) ^[1]	
- City	3,439,100
- Density	3,856.3/km ² (9,987.7/sq mi)
- Metro	5,000,000
Time zone	CET (UTC+1)
- Summer (DST)	CEST (UTC+2)

[[wiki]] This article **may need to be wikified to meet Wikipedia's quality standards**. Please [help](#) by adding *relevant internal links*, or by improving the article's *layout*. (December 2008)

Contents: [Top](#) · [0-9](#) · [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

Contents [\[hide\]](#)


- 1 Source
- 2 Chemistry
- 3 Target
- 4 Mode of action
- 5 Toxicity
- 6 Treatment
- 7 References

 The **neutrality of this section is disputed**. Please see the discussion on the [talk page](#). Please do not remove this message until the [dispute is resolved](#). (December 2008)

 This section requires [expansion](#).

 [27.964006°N 82.444239°W](#)

en This user is a **native** speaker of **English**.

 This section **may require cleanup to meet Wikipedia's quality standards**. Please [improve this section](#) if you can. (August 2010)

→ The English Wikipedia contains more than 200 000 templates.

Investigating IQ Flaws of Wikipedia Articles

Wikipedia cleanup templates

Area	
- City	891.82 km ² (344.3 sq mi)
Elevation	34 - 115 m (-343 ft)
Population (2009-09-30) ^[1]	
- City	3,439,100
- Density	3,856.3/km ² (9,987.7/sq mi)
- Metro	5,000,000
Time zone	CET (UTC+1)
- Summer (DST)	CEST (UTC+2)


[[wiki]] This article **may need to be wikified to meet Wikipedia's quality standards**. Please [help](#) by adding *relevant internal links*, or by improving the article's *layout*. (December 2008)

Contents: [Top](#) · [0-9](#) · [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

Contents [hide]


- 1 Source
- 2 Chemistry
- 3 Target
- 4 Mode of action
- 5 Toxicity
- 6 Treatment
- 7 References

 The **neutrality of this section is disputed**. Please see the discussion on the [talk page](#). Please do not remove this message until the [dispute is resolved](#). (December 2008)

 This section requires [expansion](#).

 [27.964006°N 82.444239°W](#)

en This user is a [native speaker of English](#).

 This section **may require cleanup to meet Wikipedia's quality standards**. Please [improve this section](#) if you can. (August 2010)

- 333 cleanup templates identified using an automatic retrieval approach.
- 414 642 (13%) articles containing at least one cleanup template.

Investigating IQ Flaws of Wikipedia Articles

Wikipedia cleanup templates related to concrete IQ flaws

Area	
- City	891.82 km ² (344.3 sq mi)
Elevation	34 - 115 m (-343 ft)
Population (2009-09-30) ^[1]	
- City	3,439,100
- Density	3,856.3/km ² (9,987.7/sq mi)
- Metro	5,000,000
Time zone	CET (UTC+1)
- Summer (DST)	CEST (UTC+2)

[[wiki]] This article **may need to be wikified to meet Wikipedia's quality standards**. Please [help](#) by adding *relevant internal links*, or by improving the article's *layout*. (December 2008)

Contents: [Top](#) · [0-9](#) · [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)


- Contents** [hide]
- 1 Source
 - 2 Chemistry
 - 3 Target
 - 4 Mode of action
 - 5 Toxicity
 - 6 Treatment
 - 7 References

 The **neutrality of this section is disputed**. Please see the discussion on the [talk page](#). Please do not remove this message until the [dispute is resolved](#). (December 2008)

 This section requires [expansion](#).

 27.964006°N 82.444239°W

en This user is a **native** speaker of **English**.

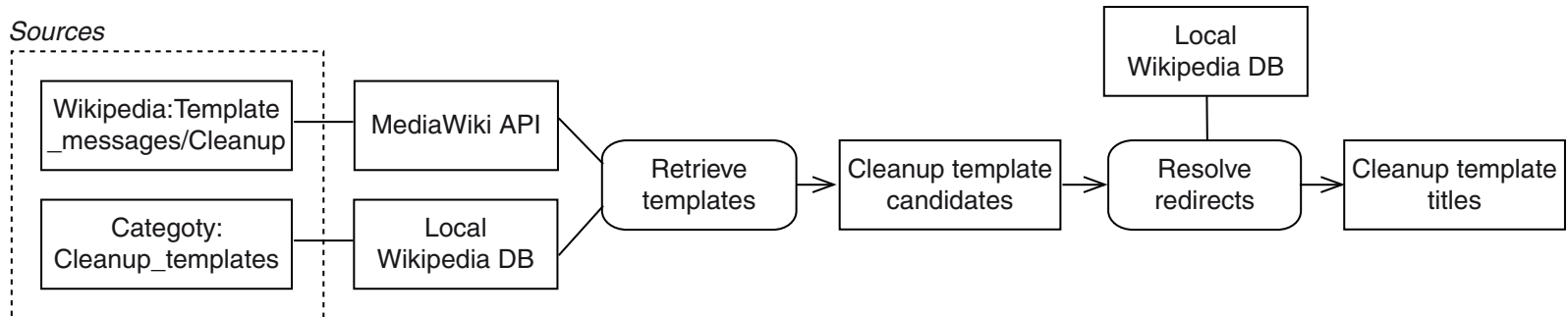
 This section **may require cleanup to meet Wikipedia's quality standards**. Please [improve this section](#) if you can. (August 2010)

→ 73 IQ flaw related cleanup templates identified by a manual analysis.

Investigating IQ Flaws of Wikipedia Articles

Cleanup template retrieval

- ❑ *Problem.* No straight forward way to make out cleanup templates.
- ❑ *Approach.* Examine meta information about cleanup templates:
 1. Meta page *Wikipedia:Template_messages/Cleanup* and
 2. Wikipedia category *Category:Cleanup_templates*.



Investigating IQ Flaws of Wikipedia Articles

Cleanup template analysis

- Check the cleanup templates against the following criteria:
 - *Scope*. Refers to the whole article.
 - *Concreteness*. Describes a single and concrete cleanup task.
 - *Generality*. Not specific to a certain domain, language, or user group.

- Cleanup templates fulfilling all criteria / IQ flaws:
 - Unreferenced
 - Refimprove
 - Orphan
 - No footnotes
 - Notability
 - Trivia
 - Original research
 - Citations missing
 - POV
 - Wikify
 - Inappropriate tone
 - Advert
 - More footnotes
 - Lead too short
 - ...

Problem Definition

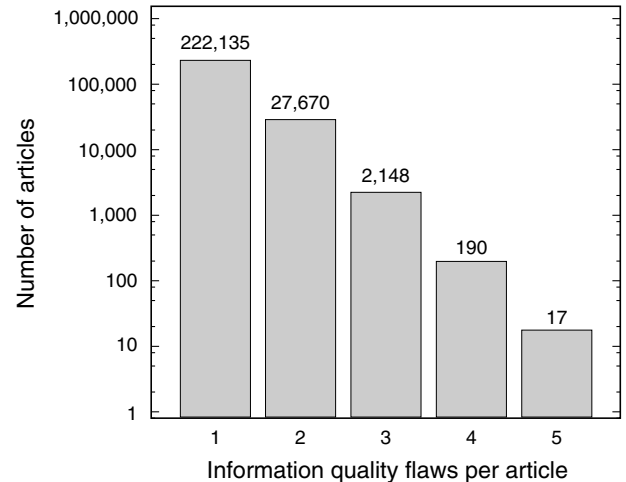
The automatic detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.

- F is the set of IQ flaws occurring in Wikipedia articles. → ✓
- D is the set of low-quality Wikipedia articles, where each $d \in D$ has at least one IQ flaw $f \in F$. → ✓
- $D_c \subset D$ is a corpus containing pre-classified articles. → ?
- $\alpha : D \rightarrow \mathbf{D}$ is an article quality model. → ?
- $c : \mathbf{D} \rightarrow F$ is a multiclass classifier. → ?

IQ Flaw Corpus

- The 73 cleanup templates serve as human labels.
- 64 of these cleanup templates actually occur in the Wikipedia snapshot.
- 223 278 articles containing exactly one of these cleanup templates.
- Multilabeled, redirect, list, and disambiguation articles are discarded.

Number of examples	Number of classes
> 100.000 (52%)	1
50.000 - 100.000	0
10.000 - 50.000 (29%)	2
1.000 - 10.000 (16%)	14
100 - 1.000 (2%)	16
< 100 (1%)	31



Problem Definition

The automatic detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.

- F is the set of IQ flaws occurring in Wikipedia articles. → ✓
- D is the set of low-quality Wikipedia articles, where each $d \in D$ has at least one IQ flaw $f \in F$. → ✓
- $D_c \subset D$ is a corpus containing pre-classified articles. → ✓
- $\alpha : D \rightarrow \mathbf{D}$ is an article quality model. → ?
- $c : \mathbf{D} \rightarrow F$ is a multiclass classifier. → ?

Article Quality Model

Features

- ❑ 40-50 article features from previous research.
- ❑ 10-15 new features.
- ❑ Classified by the source of information:

Content-based

→ plain text

- Character count
- Word count
- Syllables counts
- Readability indices
- Part of speech tags
- Passive voice count
- ...

Structural

→ wiki syntax

- Link counts
- Image count
- Link distribution
- Section sizes
- Heading structure
- References counts
- ...

History-based

→ MediaWiki API

- Currency
- Number of edits
- Editor counts
- Number of reverts
- Edits per editor
- Revert time
- ...

Problem Definition

The automatic detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.

- F is the set of IQ flaws occurring in Wikipedia articles. → ✓
- D is the set of low-quality Wikipedia articles, where each $d \in D$ has at least one IQ flaw $f \in F$. → ✓
- $D_c \subset D$ is a corpus containing pre-classified articles. → ✓
- $\alpha : D \rightarrow \mathbf{D}$ is an article quality model. → ✓
- $c : \mathbf{D} \rightarrow F$ is a multiclass classifier. → Current work

IQ Flaw Classification

One-against-all

- $|F| = 64$ binary classifiers.
- The i th classifier c_i is trained taking the examples from the i th class $f_i \in F$ as positive and the examples from all other classes as negative.
- Winner-takes-all strategy: A new example $d \in D \setminus D_c$ is assigned to the class f_i if c_i has the largest confidence value.

One-against-one

- $|F|(|F| - 1)/2 = 2016$ binary classifiers.
- The classifier c_{ij} is trained taking the examples from the i th class $f_i \in F$ as positive and the examples from the j th class $f_j \in F$ as negative.
- Max-wins voting: For a new example $d \in D \setminus D_c$ the classifier c_{ij} votes for f_i or f_j , respectively. After each classifier makes its vote, d is assigned to the class with the largest number of votes.

Summary

What we have done:

- ❑ Proposed the detection of IQ flaws in Wikipedia articles.
- ❑ Identified the IQ flaws actually occurring in Wikipedia articles.
- ❑ Human-labeled IQ flaw corpus.
- ❑ Article quality model.
- ❑ IQ flaw classification approaches.

Summary

What we have done:

- ❑ Proposed the detection of IQ flaws in Wikipedia articles.
- ❑ Identified the IQ flaws actually occurring in Wikipedia articles.
- ❑ Human-labeled IQ flaw corpus.
- ❑ Article quality model.
- ❑ IQ flaw classification approaches.

Open problems / work in progress:

- ❑ Find the best IQ flaw classification strategy.
- ❑ Evaluation.
- ❑ Combine related IQ flaws.
- ❑ Multilabel classification.

Thank you!