Automatic Detection of Information Quality Flaws in Wikipedia Articles

Maik Anderka
Bauhaus-Universität Weimar
www.webis.de

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.
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 automatically detection of IQ flaws in Wikipedia articles is addressed by means of machine learning.
Problem Definition

The automatically 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 automatically 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 automatically 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 automatically 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.
- $\alpha : D \rightarrow \mathbb{D}$ is an article quality model.
Problem Definition

The automatically 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.
- $\alpha : D \to D$ is an article quality model.
- $c : D \to F$ is a multiclass classifier.
Problem Definition

The automatically 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.
- $\alpha : D \rightarrow D$ is an article quality model.
- $c : D \rightarrow F$ is a multiclass classifier.
- $c : D \rightarrow \mathcal{P}(F)$ is a multiclass multilabel classifier.
Problem Definition

The automatically 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.
- $\alpha : D \rightarrow D$ is an article quality model.
- $c : D \rightarrow F$ is a multiclass classifier.
Problem Definition

The automatically 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.
- $\alpha : D \to \mathcal{D}$ is an article quality model.
- $c : D \to F$ is a multiclass classifier.

Previous work
Problem Definition

The automatically 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.
- $\alpha : D \rightarrow D$ is an article quality model.
- $c : D \rightarrow F$ 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.
The English Wikipedia contains more than 200,000 templates.
333 cleanup templates identified using an automatic retrieval approach.

414,642 (13%) articles containing at least one cleanup template.
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
  - ...

©www.webis.de 2010
Problem Definition

The automatically 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.  →  ?
- $\alpha : D \rightarrow D$ is an article quality model.  →  ?
- $c : D \rightarrow F$ 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.

<table>
<thead>
<tr>
<th>Number of examples</th>
<th>Number of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 100,000</td>
<td>1 (52%)</td>
</tr>
<tr>
<td>50,000 - 100,000</td>
<td>2 (29%)</td>
</tr>
<tr>
<td>10,000 - 50,000</td>
<td>14 (16%)</td>
</tr>
<tr>
<td>1,000 - 10,000</td>
<td>16 (2%)</td>
</tr>
<tr>
<td>100 - 1,000</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>&lt; 100</td>
<td>31 (1%)</td>
</tr>
</tbody>
</table>

Number of articles: 222,135
Information quality flaws per article:
- 1 flaw: 27,670
- 2 flaws: 2,148
- 3 flaws: 190
- 4 flaws: 17

©www.webis.de 2010
Problem Definition

The automatically 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.  
  \[ \rightarrow \checkmark \]

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

- $D_c \subset D$ is a corpus containing pre-classified articles.  
  \[ \rightarrow \checkmark \]

- $\alpha : D \rightarrow D$ is an article quality model.  
  \[ \rightarrow ? \]

- $c : D \rightarrow F$ is a multiclass classifier.  
  \[ \rightarrow ? \]
### 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 automatically 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. ✓
- $\alpha : D \rightarrow D$ is an article quality model. ✓
- $c : D \rightarrow F$ 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!