Towards Automatic Quality Assurance in Wikipedia

Maik Anderka Benno Stein Nedim Lipka

Bauhaus-Universität Weimar
99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

ABSTRACT

Featured articles in Wikipedia stand for high information quality, and it has been found interesting to researchers to analyze whether and how they can be distinguished from "ordinary" articles. Here we point out that article discrimination falls far short of writer support or automatic quality assurance: Featured articles are not identified, but are made. Following this motto we compile a comprehensive list of information quality flaws in Wikipedia, model them according to the latest state of the art, and devise one-class classification technology for their identification.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces— *Evaluation/methodology*

General Terms: Measurement, Algorithms, Experimentation

Keywords: Wikipedia, Information Quality, Flaw Detection

1. WIKIPEDIA AND QUALITY

Existing research on information quality in Wikipedia mostly deals with featured article identification. For this purpose articles are analyzed with respect to the number of edits and editors [10], the number of words [1], or the character trigrams distribution [6]. However, there is only little research from a constructive viewpoint, such as vandalism detection [7] or flaw identification for example, which would help writers to improve an article. Stvilia et al. [8] is one of the exceptions who take a step towards quality assurance, namely by defining computational information quality metrics for Wikipedia articles.

Our contribution to improve quality assurance is twofold. Firstly, detailed in the remainder of this section, we report on our exploratory analysis targeting the information quality flaws in articles. Secondly, outlined in Section 2, we report on flaw detection.

We extracted and analyzed cleanup template messages in the English Wikipedia, since they are often used to tag flaws in articles. From the 333 different message types we found, we consider 70 as true information quality flaws, as they all show the following three properties: they describe a single and specific information quality aspect, they refer to an article as a whole, and they are not restricted to a particular domain, language, or user group.

To guarantee reproducibility the analyses presented in this paper are based on the English Wikipedia snapshot from January 16, 2010, which is provided by the Wikimedia Foundation. Altogether, we found that from the 2 958 303 English Wikipedia articles, 252 160 articles (8.52%) have been tagged to contain at least one of the 70 flaws. The by far most frequent one is the *Unreferenced* flaw; it occurs in 135 210 articles (4.57%). An article is called unreferenced if it does not cite any references or sources. The number of flaws per article differs from one to five, whereas the majority of the tagged articles (88%) are tagged with exactly one flaw.

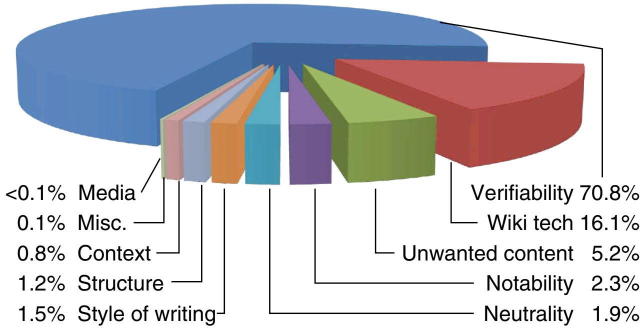

Figure 1: Our classification system of 10 flaw types and its distribution in the English Wikipedia snapshots from January 2010. The percentages relate to the set of the 252 160 tagged articles.

To break down the general information quality situation within Wikipedia, we propose 10 general flaw types and organize the 70 flaws along these types. Figure 1 lists these types and shows their distribution within the snapshot. The majority of the flaws concerns an article's verifiability, which is one of the most important principles of an encyclopedia. Note that at least 6.66% of all English Wikipedia articles are tagged with some verifiability flaw. Due to the size and the few control mechanisms in Wikipedia, it is more then likely that many flawed articles are not yet identified. I.e., our analysis underestimates the actual frequencies.

2. FLAW DETECTION

Let D be the set of Wikipedia articles. We model the quality-specific characteristics of an article $d \in D$ as document vector $\mathbf{d}$, where each dimension in $\mathbf{d}$ quantifies one of altogether 88 quality assessment features. Our document model combines state of the art features from the relevant literature with efficient new features that quantify the usage of in-links, templates, lists, and special words, among others.

We interpret the detection of a flaw f as a one-class classification problem—presuming that only information about one class, the so-called target class, is available. Here, the target class of some flaw f is made up of all articles that are tagged with f. For an in depth discussion of one-class classification see Tax [9].

For each flaw a specific one-class classifier c, $c : \mathbf{D} \to \{1, 0\}$

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

Table 1: Performance for each of the five most frequent flaws. The flaws are organized along the frequency classes F_1, F_2, and F_3. Due to its construction the classifier performs with a stable recall of 0.9.

	Flaw name	Precision	Recall	F-measure	AUC
F_1	Unreferenced	0.96	0.9	0.93	0.95
F_2	Refimprove	0.82		0.86	0.90
	Orphan	1.00		0.94	0.97
F_3	No footnotes	0.92		0.91	0.93
	Notability	0.99		0.94	0.96

is trained, based on the articles of the respective target class; $\mathbf{D}$ denotes the set of document vectors for the articles in D. We resort to a one-class classification approach, which combines density estimation with class probability estimation [4]. We apply bagged random forest classifiers with 1 000 decision trees [5] as class probability estimator, using 10 bagging iterations. The approach calculates the absolute probability $P(\mathbf{d} \mid f)$, whereas an article d is classified as flawed by f if $P(\mathbf{d} \mid f) > \tau$. The threshold τ is derived empirically from the target rejection rate of the classifier, which is the rejection rate of the target class in the training phase. We adjust the target rejection rate to 0.1, which leads to an classifier effectiveness of about 0.9 in terms of recall.

Experiment Setup In order to evaluate c with respect to its precision one needs a representative sample of examples from outside the target class, so-called outliers. Typically, a one-class classifier is evaluated on generated examples, assuming that outliers are uniformly distributed [9]. Here, we use a set of 1 000 randomly sampled featured articles as outliers. This rather optimistic approach is based on the hypothesis that featured articles do not contain an information quality flaw at all. Note that we knowingly accept a systematic bias since featured articles cannot be considered as a representative sample of flawless Wikipedia articles.

For each flaw f, its one-class classifier c is evaluated with 1 000 articles, which are randomly sampled from the respective target class (articles tagged with f) and the 1 000 outliers (featured articles), using ten-fold cross-validation. Within each run the classifier is trained with 900 articles from the target class; testing is performed with the remaining 100 articles plus 100 outliers. Note that c is trained exclusively with the examples of the respective target class, i.e., c is neither affected by the class distribution nor by the featured articles.

Effectiveness of Detecting Flaws Table 1 shows the performance values for the five most frequent flaws. For all practical purposes, precision is the determining measure of effectiveness; consider for instance a bot that autonomously tags flawed articles. The areas under the ROC curves (AUC) [3] with values of greater than 0.5 indicate that the predictions are not based on random guessing. Certain flaws can be detected with a nearly perfect precision, e.g., *Orphan* and *Notability*. For other flaws the precision deteriorates significantly, e.g., *Refimprove*. Two possible explanations are the following: (1) The inability of the document model to capture the gist of certain flaws. (2) An inappropriate one-class classification approach for the particular problem.

Expected Performance in the Wild Based on the recall and the false positive rate of a classifier for the balanced test set, we compute the precision for varying class distributions 1:n (the ratio of articles containing a flaw f and articles not containing f). Figure 2 shows precision values as a function of the flaw distribution. To assess the performance that can be expected in the wild, we organize the flaws along the three frequency classes F_1, F_2, and F_3. Observe that the expected precision values (the highlighted portions of the curves) for the flaws *Unreferenced*, *Orphan*, and *Notability* are still high. Although the flaw *Notability* belongs to frequency class F_3, the expected precision is still about 0.9, which shows that the one-class classifier captures the concept of this flaw exceptionally well. The expected precision values for the two remaining flaws are low (about 0.5 or less). Aside from the conceptual weaknesses mentioned above, this might also be an indication for the fact that the training set of the classifiers is too small.

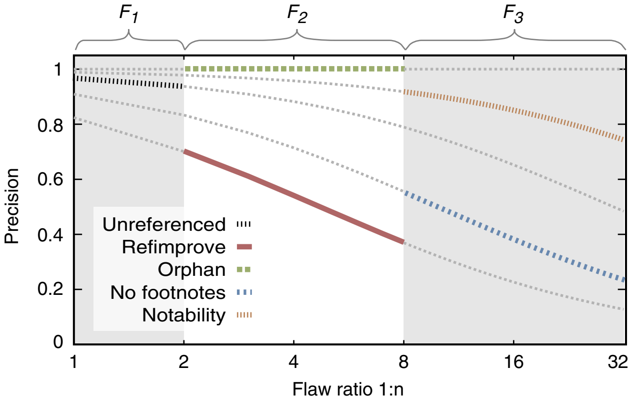

Figure 2: Precision over flaw ratio for each of the five most frequent flaws: 1:n ∼ flawed:flawless, with n $\in [1, 32]$.

3. RESEARCH OUTLOOK

Our current research targets the development of tailored one-class classifiers: (1) Instead of resorting to a single document model, a flaw-specific view is developed, which combines expert rules, multi-level filtering, and feature selection. (2) In addition, the positive and negative impacts of different one-class classification approaches (density estimation, boundary identification, reconstruction analysis) on the flaw-specific document models are investigated.

Based on the lessons learned, we plan to operationalize our classification approach as a Wikipedia bot that tags articles autonomously. This will also support the principle of *intelligent task routing* [2], which addresses the automatic delegation of particular flaws to appropriate human correctors.

4. REFERENCES

[1] J. Blumenstock. Size matters: word count as a measure of quality on Wikipedia. In *Proceedings of WWW'08*.

[2] D. Cosley, D. Frankowski, L. Terveen, and J. Riedl. Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In *Proceedings of SIGCHI'06*.

[3] T. Fawcett. ROC graphs: notes and practical considerations for researchers. Technical report, HP Laboratories, 2004.

[4] K. Hempstalk, E. Frank, and I. Witten. One-class classification by combining density and class probability estimation. In *Proceedings of ECML PKDD'08*.

[5] T. K. Ho. Random decision forests. In *Proceedings of ICDAR'95*.

[6] N. Lipka and B. Stein. Identifying featured articles in Wikipedia: writing style matters. In *Proceedings of WWW'10*.

[7] M. Potthast, B. Stein, and R. Gerling. Automatic vandalism detection in Wikipedia. In *Proceedings of ECIR'08*.

[8] B. Stvilia, M. Twidale, L. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *Proceedings of ICIQ'05*.

[9] D. Tax. *One-class classification*. PhD thesis, Technische Universiteit Delft, 2001.

[10] D. Wilkinson and B. Huberman. Cooperation and quality in Wikipedia. In *Proceedings of WikiSym'07*.