Automatically Assessing the Quality of Wikipedia Articles

By
Joshua E. Blumenstock (jblumenstock@berkeley.edu)
School of Information, UC Berkeley

UCB iSchool Report 2008-021
April 2008

Abstract:
Since its inception in 2001, Wikipedia has fast become one of the Internet's most dominant sources of information. Dubbed "the free encyclopedia", Wikipedia contains millions of articles that are written, edited, and maintained by volunteers. Due in part to the open, collaborative process by which content is generated, many have questioned the reliability of these articles. The high variance in quality between articles is a potential source of confusion that likely leaves many visitors unable to distinguish between good articles and bad. In this work, we describe how a very simple metric – word count – can be used to as a proxy for article quality, and discuss the implications of this result for Wikipedia in particular, and quality assessment in general.

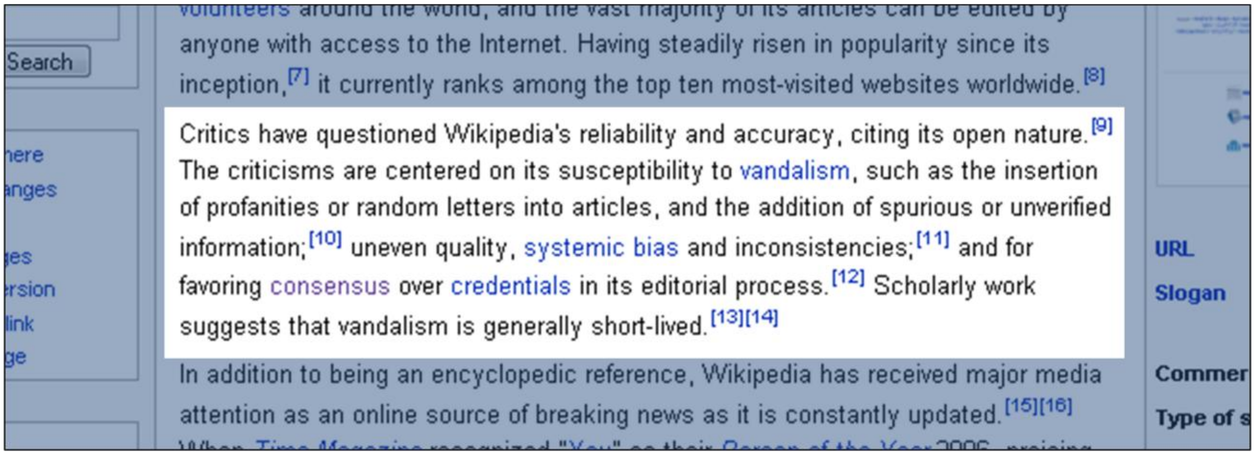
Keywords:
Wikipedia, information quality, user-generated content, word count, quality

1 Introduction

1.1 Background on Wikipedia

Since its inception in 2001, Wikipedia, "the free encyclopedia", has quickly become a dominant source of information on the Internet. As of March 2008, it is the ninth most popular website worldwide.[1] Globally, there are over 9.25 million articles in 253 languages, with over 2.3 million articles in the English version alone. Perhaps the most notable feature of Wikipedia's popularity is the fact that the articles are written by unpaid users, rather than paid experts. Anyone can visit Wikipedia.org and create a new article, or modify an existing one. To date, over 200 million such edits have been performed by visitors to Wikipedia.[2]

The fact that Wikipedia articles are created in such an organic fashion has caused many to question the quality of Wikipedia articles. Some articles lack essential information, some contain factual inaccuracies, and some are just poorly written. As author Nicholas Carr decried, "this is garbage, an incoherent hodge-podge of dubious factoids that adds up to something far less than the sum of its parts…"[3] A more somber assessment can be found on the pages of Wikipedia itself (Figure 1):



volunteers around the world, and the vast majority of its articles can be edited by anyone with access to the Internet. Having steadily risen in popularity since its inception,[7] it currently ranks among the top ten most-visited websites worldwide.[8]

Critics have questioned Wikipedia's reliability and accuracy, citing its open nature.[9] The criticisms are centered on its susceptibility to vandalism, such as the insertion of profanities or random letters into articles, and the addition of spurious or unverified information;[10] uneven quality, systemic bias and inconsistencies;[11] and for favoring consensus over credentials in its editorial process.[12] Scholarly work suggests that vandalism is generally short-lived.[13][14]

In addition to being an encyclopedic reference, Wikipedia has received major media attention as an online source of breaking news as it is constantly updated.[15][16]

Figure 1. Excerpt from Wikipedia describing the quality of Wikipedia articles.[4]

Such criticism notwithstanding, Wikipedia is thought by many to be a high quality source of information, and a number of recent studies seem to support this point of view. A recent study published in *Nature*, for instance, compared Wikipedia articles to articles from Encyclopedia Britannica, and found the two to be of comparable quality. Among other things, the study noted that, "the exercise revealed numerous

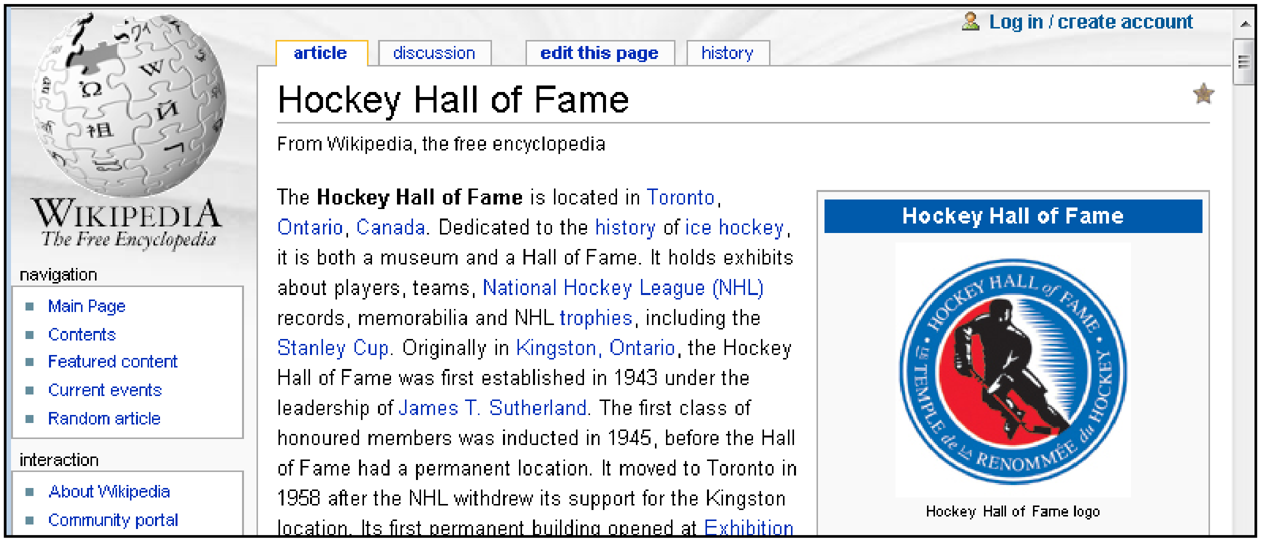
[1] http://alexa.com/site/ds/top_sites?ts_mode=global&lang=none
[2] http://en.wikipedia.org/wiki/Special:Statistics
[3] "Wikipedia founder admits to serious quality problems", The Register, Tuesday 18th October 2005.
[4] http://en.wikipedia.org/wiki/Wikipedia

errors in both encyclopaedias, but among 42 entries tested, the difference in accuracy was not particularly great: the average science entry in Wikipedia contained around four inaccuracies; Britannica, about three."[5]

1.2 Featured Articles

The high variance in quality of Wikipedia articles makes it difficult for many visitors to trust the content they encounter. To highlight content of exceptional quality, Wikipedia highlights articles of exceptional quality by making them "featured" articles. These articles are to be trusted as reliable and accurate. As Wikipedia.org explains, "featured content represents the best that Wikipedia has to offer. These are the articles, pictures, and other contributions that showcase the polished result of the collaborative efforts that drive Wikipedia. All featured content undergoes a thorough review process to ensure that it meets the highest standards and can serve as an example of our end goals."[6] The review process involves careful scrutiny by the community of editors as well as selection by a designated director. [7] See Figure 2 for an example of a featured article.

Unfortunately, only one article in every thousand articles is featured. While some articles are marked with similar metadata to denote exceptionally low quality, the vast majority of articles contain no such markings, and the visitor is left to figure out for herself whether the article should be trusted. It is this uncertainty that has motivated researchers to seek a means of easily measuring article quality.



Hockey Hall of Fame

From Wikipedia, the free encyclopedia

The **Hockey Hall of Fame** is located in Toronto, Ontario, Canada. Dedicated to the history of ice hockey, it is both a museum and a Hall of Fame. It holds exhibits about players, teams, National Hockey League (NHL) records, memorabilia and NHL trophies, including the Stanley Cup. Originally in Kingston, Ontario, the Hockey Hall of Fame was first established in 1943 under the leadership of James T. Sutherland. The first class of honoured members was inducted in 1945, before the Hall of Fame had a permanent location. It moved to Toronto in 1958 after the NHL withdrew its support for the Kingston location. Its first permanent building opened at Exhibition

Figure 2. Example featured article. The star in the top-right corner denotes the featured status.

[5] *Nature News* **438**, 900-901 (15 December 2005)
[6] http://en.wikipedia.org/wiki/Portal:Featured_content
[7] http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

2 Related Work

2.1 Qualitative Work

Beyond the editorial guidelines found at Wikipedia.org, there has been substantial qualitative work aimed at understanding quality in encyclopedias in general and Wikipedia in particular. Crawford (2001), for instance, presents a thorough framework for assessing encyclopedia quality. Lih (2003) proposed further metrics for the online context, and analyzed the correlation between the numbers of revisions and unique authors to Wikipedia articles and the quality of these articles. Heuristically, Lih (2003) proposed using the total number of edits and unique editors to measure article quality and Cross (2006) suggested coloring text according to its age, to give readers some indication of quality.

2.2 Quantitative Work

Other researchers have designed far more complex systems for measuring article quality. This work typically relies on techniques from machine learning, and seeks to produce algorithmic methods of measurement. The standard methodology roughly involves three steps:

i. **Feature extraction.** This involves representing each article as a combination of different quantifiable metrics. The metrics, called features, might include straightforward information such as *sentence count*, *word count*, *syllable count*; more structural features such as *image count*, *number of references*, *number of links*; linguistic information like *number of noun phrases, ratio of verbs/adverbs*; revision history such as *edit count*, *number of unique editors*, or derived information such as the gunning fog index, the Fleisch-Kincaid readability measure, etc.

ii. **Classification/Quality prediction.** Algorithms are used to predict the quality of an article based on its features. For instance, if we believed the age of the article to be its most important feature, we might predict that old articles would be of high quality and new articles would be of low quality.

iii. **Evaluation.** The predicted quality is measured against an objective standard of quality. While a few studies, such as recent work by Adler and de Alfaro (2007), have used human experts to judge the quality of the predictions, the most common approach is to use featured articles as a proxy for quality. Thus, the algorithm in step (ii) will try to correctly classify each article as a featured article, or a not featured article, and the accuracy is measured as the number of correct classifications divided by the total number of articles classified.

A primary advantage of this methodology is that it is objective and automatic. Once an effective measure of quality is found it can, in principle, be applied to any article on Wikipedia.

Following this approach, Stvilia et al. (2005) tried to devise a system that would compute the quality of an article, based on the qualitative standards for quality described in Crawford (2001). Crawford named seven factors important to quality: (1) Scope (2) Format; (3) Uniqueness; (4) Authority; (5) Accuracy; (6)

Currency; and (7) Accessibility. Stivilia et al. transformed these factors into quantifiable composites: (1) *Consistency* = 0.6*Admin. Edit Share + 0.5*Age; (2) *Authority/Reputation* = 0.2*Unique Editors + 0.2*Total Edits + 0.1*Connectivity + 0.3* Reverts + 0.2* External Links + 0.1*Registered User Edits + 0.2* Anonymous User Edits; etc. Having computed these factors for each article, Stvilia et al. then used cluster analysis to predict whether each article was featured or not, and achieved 86% overall accuracy.

Related work by Zeng et al. (2006) attempted to measure the "trust" of an article based on its edit history. Here, the relevant features were metrics such as the number of revisions, the number of deletions, and the number of blocked authors editing each article. Based on these features, Zeng et al. used a dynamic Bayesian network to model the evolution of each article, and found that they could classify featured articles with 84% accuracy.

In contrast to the complex methods described above, Blumenstock (2008) detailed how word count can be used to identify featured articles with greater than 97% accuracy. It is this result, and its potential applications, that are discussed below.

3 Methods

3.1 Building a corpus of articles

To build a corpus of data, we started with the roughly 5.6 million articles from the July 2007 version of the English Wikipedia. We removed all articles with fewer than fifty words, as well as all non-textual files such as images, templates and lists. After computing structural features (see section 3.2 below), all Wikipedia-related markup and formatting was removed, as was irregular content such as tables and metadata. There remained 1,554 featured articles in this cleaned dataset, and roughly a million articles that were not featured. We randomly selected 9,513 of these non-featured articles to serve as a "random" corpus in training and testing the classification algorithms.

3.2 Feature extraction

Using various custom and open-source scripts, we extracted over 100 features for each article. These features are summarized in Table 1, and include readability metrics, syntactic features, structural features, and part of speech tags. Notably absent from the features are metrics related to the revision history of each article. These features were omitted for practical reasons, as the sheer size of each snapshot (roughly 7 gigabytes) would have made it quite difficult to store and process different versions from different dates. However, as is discussed in the sections 4 and 5, it is not clear that such information would improve the accuracy of classification.

Surface features	Structural features	Readability Metrics[8]	Part of Speech tags[9]
Characters Words Sentences Syllables Tokens One-syllable words Complex Words	Internal Links External Links Categories Images Citations Sections Tables Infoboxes … and others (27 total)	Gunning Fog Coleman-Liau Index Flesch-Kincaid Reading Ease SMOG index Automated Readability Index FORCAST readability	Noun phrases Determiners Past participle verbs Adjectives Nouns Proper Nouns Preterites Adverbs …and others (53 total)

Table 1. Features extracted for each article.

3.3 Classification

Having obtained these features, we tested a variety of classification schemes, ranging from the simple (e.g. threshold functions, regression) to the complex (e.g. random forest, multi-layer perceptron).[10] For all experiments, we used two thirds of the articles for training (7,378 articles) and one third for testing (3,689 articles). There was a similar ratio of featured articles to random articles in both the training and testing corpuses (roughly .16).

4 Results

We found that, using article length alone, we could correctly distinguish between featured and non-featured articles with greater than 97 percent accuracy (see Table 2). The maximum accuracy was produced with a multi-layer perceptron model, but similar results were achieved with much simpler methods. For instance, the simple heuristic of classifying all articles with more than 2,000 words as "featured" and those with fewer than 2,000 words a "random", one can achieve an accuracy of 96.31 percent. Using 1,830 words as the cutoff threshold, that accuracy increases to 96.46 percent (see Figure 3). Compared to the baseline accuracies of 86 percent (Stvilia et al.) and 84 percent (Zeng et al.), the simplicity of the word count threshold is quite compelling.

Class	n	TP rate	FP rate	Precision	Recall	F-measure
Featured	1554	.936	.023	.871	.936	.902
Random	9513	.977	.064	.989	.977	.983

Table 2. Classifier performance. Using word count as input data, multi-layer perceptron model.

[8] Computed using tools available at http://www.gdssw.com/tools/

[9] POS tagging was done with the GATE package, available at http://www.gate.ac.uk/

[10] Algorithms were implemented in R (http://www.r-project.org/), WEKA (http://www.cs.waikato.ac.nz/ml/weka/), and the 'Bow' toolkit (http://www.cs.cmu.edu/~mccallum/bow/).

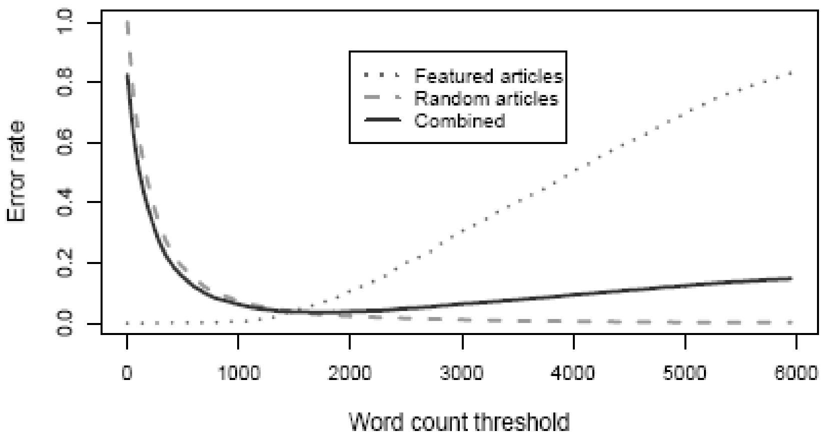


Figure 3. Effect of word count threshold on error rate. The overall error rate is minimized at a cutoff threshold of 1,830 words. However, small changes in the threshold (+/-300 words) do not significantly affect the error rate.[11]

It was possible to improve upon the 97 percent accuracy of the word count classifier by including other features from Table 1. However, the improvements in accuracy were modest: even using the full "kitchen sink" of 100 features, the maximum accuracy achieved by any classifier was 97.99 percent. For a more detailed comparison of the performance of the different classification algorithms, see Blumenstock (2008).

5 Discussion

The results from Blumenstock (2008) show that word count alone can differentiate featured Wikipedia articles from normal Wikipedia articles. In some senses, this result is intuitive: featured articles are long, and long articles are featured (see Figure 4). What is more remarkable is just how well word count works as a classifier. Not only does word count beat the other individual metrics that we expected to be effective discriminators (such as number of references, readability metrics, etc.), it significantly outperforms complex techniques based on revision history, and performs nearly as well as a classifier utilizing all one hundred metrics.

The fact that featured articles can be so reliably identified based on their length calls into question the common practice of using featured articles as a proxy for quality. For researchers seeking to computationally measure the quality of articles, a more robust gold standard of quality is needed. Human panels are expensive and subjective, but bottom-up efforts to generate quality ratings, such as the Wikiproject Biography/Assessment,[12] offer a promising source for future research.

These results also draw our attention to the counterexamples – long articles that haven't been featured, and short articles that have. For instance, we might ask whether all long Wikipedia articles are of high quality, and therefore more likely to be featured. Such a finding would be natural given the collaborative

[11] Figure excerpted from Blumenstock (2008), with permission.
[12] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Biography/Assessment

process through which articles are created. As articles grow, they likely receive the attention of more editors, and thus the quality would be expected to improve. Similarly, we might ask whether it is possible for short articles to be of high quality. In other contexts it is clearly possible for short content to be of high quality; in the Wikipedia context it might not be easy for a high quality article to remain short – the attention it receives might spoil its brevity.

In our dataset, we can superficially address these questions by looking at articles that were misclassified by the word count heuristic. These misclassifications fell into two categories: (i) featured articles that were incorrectly predicted to be random, and (ii) random articles that were incorrectly labeled as featured. Some examples from the first category include "Music of Ireland" and "African-American Civil Rights Movement (1955-1968)"; examples from the second include "Victoria Cross for New Zealand", "Stevie Nicks", and "Bertrand Russell". To our untrained eyes, it was not obvious that the articles in the former category were of lesser quality than those in the latter. In other words, "Music of Ireland" seemed to be of comparable quality to "Bertrand Russell," even though Wikipedia's editors say otherwise. People more familiar with Wikipedia's quality guidelines could likely see the difference, but it would be quite surprising if they could beat the 97% accuracy threshold achieved by the word count classifier.

Featured articles are meant to be "the best that Wikipedia has to offer"; these results indicate that they might merely be the longest Wikipedia has to offer. The high degree to which word count can approximate Wikipedia's elaborate peer-review process is somewhat unsettling. However, this tight symmetry is not necessarily undesirable; as noted above, it is possible that the collaborative nature of Wikipedia forces long articles to be of a high quality. In the end, whether one views these results as an expected correlation or an indication of something more insidious is more a matter of perspective than a question to be answered by academic research.

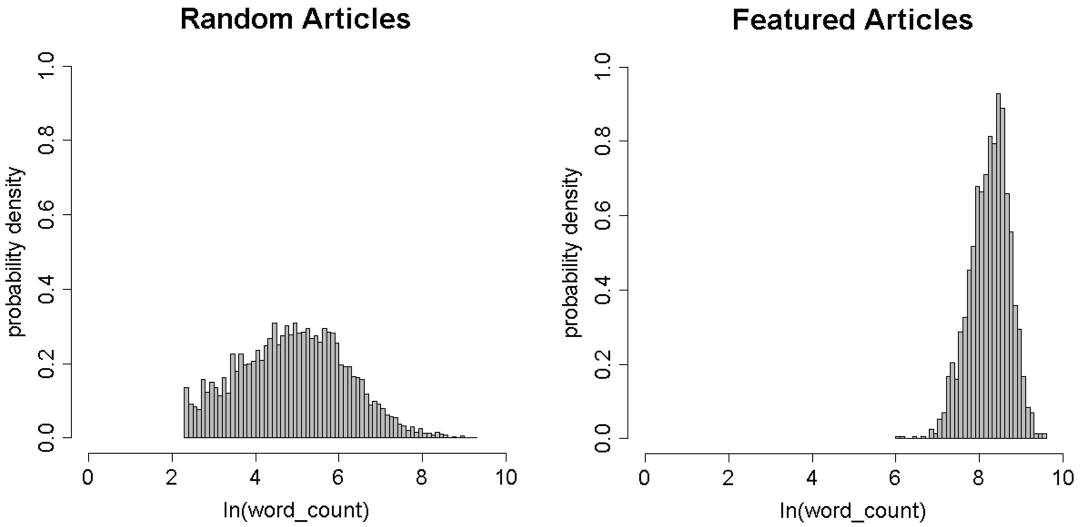


Figure 4. Distribution of word count for featured and random articles.[13]

[13] Figure excerpted from Blumenstock (2008), with permission.

6 Conclusion

We have described how word count can be used to predict, to a very high degree of accuracy, whether a Wikipedia article will be featured. We do not, however, want to exaggerate its practical utility. Word count is a very crude metric that would likely not generalize to other domains and one which would, if internalized by authors, lead to poor quality indeed. At the same time, in the context of Wikipedia, it is simple, scalable, and remarkably accurate. Visitors unfamiliar with Wikipedia could do far worse than judging an article by its length.

7 References

B. T. Adler, L. de Alfaro. A content-driven reputation system for the wikipedia. *Proceedings of the 16th international conference on World Wide Web*, 2007.

J. E. Blumenstock. Size matters: Word count as a measure of quality on Wikipedia. *Proceedings of the 17th international conference on World Wide Web*, 2008 (forthcoming).

H. Crawford. Encyclopedias. In: R. Bopp, L. C. Smith (Eds.), Reference and information services: an introduction (3 ed.). (pp. 433-459). Englewood, CO: Libraries Unlimited. 2001.

Lih, A. Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. *Proceedings of the 5th International Symposiumon Online Journalism*, 2004.

D. McGuinness, H. Zeng, P. da Silva, L. Ding, D. Narayanan, and M. Bhaowal. Investigation into trust for collaborative information repositories: A Wikipedia case study. *Proceedings of the Workshop on Models of Trust for the Web*, 2006.

Stvilia, B., Twidale, M. B., Smith, L. C., Gasser, L.. Assessing information quality of a community-based encyclopedia. In: *Proceedings of the International Conference on Information Quality - ICIQ 2005,* 2005a.

Stvilia, B., M. Twidale, L. Gasser, and L. Smith. Information Quality Discussions in Wikipedia. *Proceedings of the 2005 International Conference on Knowledge Management*, 2005b.

Zeng, H., M. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness. Computing trust from revision history. *Intl. Conf. on Privacy, Security and Trust*, 2006.