Size Matters: Word Count as a Measure of Quality on Wikipedia

Joshua E. Blumenstock
School of Information
University of California at Berkeley
jblumenstock@berkeley.edu

ABSTRACT

Wikipedia, "the free encyclopedia", now contains over two million English articles, and is widely regarded as a high-quality, authoritative encyclopedia. Some Wikipedia articles, however, are of questionable quality, and it is not always apparent to the visitor which articles are good and which are bad. We propose a simple metric – word count – for measuring article quality. In spite of its striking simplicity, we show that this metric significantly outperforms the more complex methods described in related work.

Categories and Subject Descriptors: H.5.3 [Information Interfaces]: Group and Organization Interfaces– *Collaborative computing*; I.2.7 [Artificial Intelligence]: Natural Language Processing– *Text analysis*.

General Terms: Measurement, algorithms.

Keywords: Wikipedia, information quality, word count.

1. INTRODUCTION

As user-generated content grows in prominence, many web sites have employed complex mechanisms to help visitors identify high quality content. Wikipedia maintains a list of "featured" articles that serve as exemplars of good user-generated content. For an article to be featured, it must survive a rigorous nomination and peer-review process; only one article in every thousand makes the cut. Unfortunately, this is a laborious process, and many worthy articles never get the official stamp of approval. Thus, it might be useful to have an automatic means for detecting articles of unusually high quality.

A substantial amount of work has been done to automatically evaluate the quality of Wikipedia articles. At a qualitative level, Lih [3] proposed using the total number of edits and unique editors to measure article quality, and Cross [2] suggested coloring text according to age so that visitors could immediately discern its quality. Both studies, however, were primarily qualitative, and did not measure the discriminative value of such heuristics.

The quantitative approaches found in related work tend to be quite complex. Zeng et al. [6], for instance, used a dynamic bayesian network to develop a measure of trust and quality based on the edit history of an article. Adler and de Alfaro [1] devised similar metrics to quantify the reputation of authors and editors. In the work closest to our own, Stvilia et al. [5], computed 19 quantitative metrics, then used factor analysis and k-means clustering to differentiate featured from random articles.

2. METHODS

In contrast to the complex quantitative methods found in related work, we propose a much simpler measure of quality for Wikipedia articles: the length of the article, measured in words. While there are many limitations to such a metric, there is good reason to believe that this metric will be correlated to quality (see Figure 1). The simplicity of this metric presents several advantages:

- article length is easy to measure;
- many of the approaches mentioned in section 1 require information that is not easily obtained (such as the revision and history used in [6], [4], and [1]);
- other approaches typically operate in a black box fashion, with arcane parameters and results that are not easily interpreted by the average visitor to Wikipedia;
- article length performs significantly better than other, more complex methods.

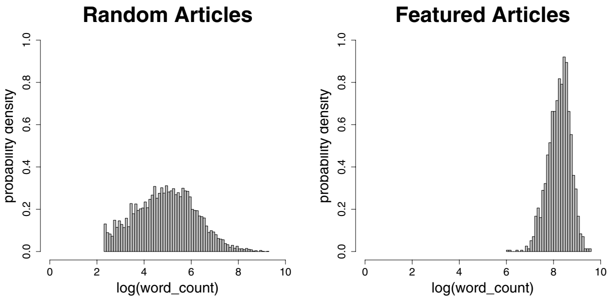

Figure 1: Word counts for featured/random articles

To test the performance of article length as a discriminant between high and low quality articles, we followed the approach taken by Zeng et al. [6] and Stvilia et al. [5] That is, instead of comparing our metric against a scalar measure of article quality, we assume that featured articles are of much higher quality than random articles, and recast the problem as a classification task. The goal is thus to maximize precision and recall of featured and non-featured articles.

To build a corpus, we extracted the full 5,654,236 articles from the 7/28/2007 dump of the English Wikipedia.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

class	n	TP rate	FP rate	Precision	Recall	F-measure
Featured	1554	0.936	0.023	0.871	0.936	0.902
Random	9513	0.977	0.064	0.989	0.977	0.983

Table 1: Performance of word count in classifying featured vs. random articles.

After stripping all Wikipedia-related markup, we removed specialized files (such as images and templates) and articles containing fewer than fifty words. This cleaned dataset contained 1,554 articles classified as "featured"; we randomly selected an additional 9,513 cleaned articles to serve as a non-featured "random" corpus. Our corpus thus contained a total of 11,067 articles. In the experiments described below, we used 2/3 of the articles for training (7,378 articles) and 1/3 for testing (3,689 articles), with a similar ratio of featured/random articles in each set.

3. RESULTS

By classifying articles with greater than 2,000 words as "featured" and those with fewer than 2,000 words as "random," we achieved 96.31% accuracy in the binary classification task.[1] The threshold was found by minimizing the error rate on the training set (see Figure 2). The reported accuracy results from testing on the held-out test set.

Modest improvements could be produced by more sophisticated classification techniques. A multi-layer perceptron, for instance, achieved an overall accuracy of 97.15%, with an f-measure of .902 for featured articles and .983 for random articles (see Table 1). Similar results were replicated with a k-nearest neighbor classifier (96.94% accuracy), a logit model (96.74% accuracy), and a random-forest classifier (95.80% accuracy). All techniques represent a significant improvement over the more complex methods in Stvilia et al. [4] and Zeng et al. [6], which produced 84% and 86% accuracy, respectively.

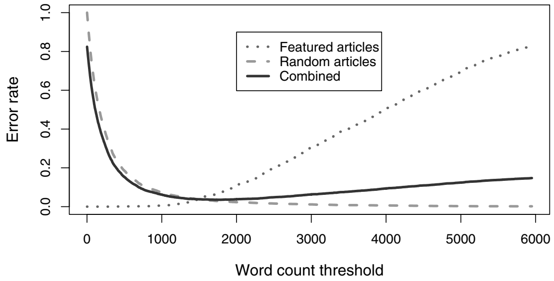

Figure 2: Accuracy of different thresholds

Given the high accuracy of the word count metric, we naturally wondered whether other simple metrics might increase classification accuracy. In other contexts, features such as part of speech tags, readability metrics, and n-gram bag-of-words have been moderately successful. In the context of Wikipedia quality, however, we found that word count was hard to beat. N-gram bag-of-words classification, for instance, produced a maximum of 81% accuracy (tested with n=1,2,3 on both svm and bayesian classifiers). Even using a "kitchen sink" of thirty features such as those listed in Table 2, no classifier achieved greater than 97.99% accuracy – a modest improvement given the considerable effort required to produce these metrics and build the classifiers.

Frequency counts		
character count	complex word count	sentence count
token count	one-syll. word count	total syllables
Readability indices		
Gunning fog index	FORCAST formula	Flesch-Kincaid
Coleman-Liau	Automatic Readability	SMOG index
Structural features		
internal links	external links	reference links
category count	image count	reference count
citation count	table count	section count

Table 2: Features from "kitchen sink" classification

4. DISCUSSION AND CONCLUSIONS

We have shown that article length is a very good predictor of whether an article will be featured on Wikipedia. Word count is a simple metric that is considerably more accurate than the complex methods proposed in related work, and performs well independent of classification algorithm and parameters.

We do not, however, mean to exaggerate the importance of this metric. By assuming that "featured" status is an accurate proxy for quality, we have implied that quality can be measured via article length. However, if our assumption does not hold, then we can only conclude that long articles are featured, and featured articles are long. Future work will explore alternative standards for quality on Wikipedia.

Acknowledgments

We thank Marti Hearst for guidance and feedback.

5. REFERENCES

[1] B. T. Adler and L. de Alfaro. A content-driven reputation system for the wikipedia. *Proc. 16th Intl. Conf. on the World Wide Web*, pages 261–270, 2007.

[2] T. Cross. Puppy smoothies: Improving the reliability of open, collaborative wikis. *First Monday*, 11, 2006.

[3] A. Lih. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. *13th Asian Media Information and Communications Centre Annual Conference*, 2004.

[4] B. Stvilia, M. Twidale, L. Gasser, and L. Smith. Information quality discussions in wikipedia. *Proc. 2005 ICKM*, pages 101–113, 2005.

[5] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. *Proc. ICIQ*, pages 442–454, 2005.

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

[1] A slightly higher accuracy of 96.46% was achieved with a threshold of 1,830 words.