Tell Me More: An Actionable Quality Model for Wikipedia

Morten Warncke-Wang
GroupLens Research
Dept. of Comp. Sci. and Eng.
University of Minnesota
Minneapolis, MN 55455, USA
morten@cs.umn.edu

Dan Cosley
Department of Info. Science
Cornell University
Ithaca, NY 14850, USA
drc44@cornell.edu

John Riedl
GroupLens Research
Dept. of Comp. Sci. and Eng.
University of Minnesota
Minneapolis, MN 55455, USA
riedl@cs.umn.edu

ABSTRACT

In this paper we address the problem of developing actionable quality models for Wikipedia, models whose features directly suggest strategies for improving the quality of a given article. We first survey the literature in order to understand the notion of article quality in the context of Wikipedia and existing approaches to automatically assess article quality. We then develop classification models with varying combinations of more or less actionable features, and find that a model that only contains clearly actionable features delivers solid performance. Lastly we discuss the implications of these results in terms of how they can help improve the quality of articles across Wikipedia.

Categories and Subject Descriptors

H.5 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Collaborative computing, Computer-supported cooperative work, Web-based interaction*

Keywords

Wikipedia, Information Quality, Modelling, Classification, Flaw Detection, Machine Learning

1. INTRODUCTION

With four million articles in the English language edition and over 25 million articles across 285 languages, Wikipedia has succeeded at building a compendium with a large number of articles. At Wikimania in 2006, Jimmy Wales held an opening plenary where he suggested Wikipedia contributors should shift their focus from the number of articles and instead work on improving their quality[1].

WikiSym '13, Aug 05-07 2013, Hong Kong, China

This raises the questions of what it means for a Wikipedia article to have quality and what actions are needed to improve it. We introduce these topics by surveying the research literature, first to understand the notion of quality and then to describe Wikipedia's quality assessment procedures, as well as the improvement actions these procedures recommend. We follow most of the prior literature by focusing on English Wikipedia; many of the concepts would work similarly in other language editions, though the details would be different.

1.1 Article quality in Wikipedia

The notion of article quality in English Wikipedia is similar to that of traditional encyclopaedias, as discussed by Stvilia et al. [29]. Accuracy in particular has received research attention, aiming to understand how a community process with no central oversight could result in articles with accuracy comparable to traditional encyclopedias [6, 14].

The notion of article quality in Wikipedia differs across both space and time. There are currently 285 language editions, each with their own user community, and research has shown that the notion of quality differs between these communities [28]. Featured Articles are considered to be the best articles Wikipedia has to offer. When they first appeared around April 2002 the only information quality criterion listed was "brilliant prose" [31]. Today, they go through a peer review process [34] that checks the articles according to "accuracy, neutrality, completeness, and style."[2] This change over time is reflected in 1,013 articles having lost their Featured Article status as of March 2013[3].

Wikipedia has other quality classes in addition to Featured Articles. There are seven *assessment classes* spanning all levels of quality from high to low: Featured Articles, Good Articles, A-, B-, C-, Start-, and Stub-class[4]. Each of these classes have specific criteria[5], for instance how completely the article covers the subject area. Out of the seven classes only the Featured Articles are regarded as mostly finished; the other six classes come with general descriptions of how improvements can be made. For example, C-class arti-

[1] http://wikimania2006.wikimedia.org/wiki/Opening_Plenary_(transcript)#Quality_initiative_.2833:20.29

[2] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles

[3] http://en.wikipedia.org/wiki/Wikipedia:Former_featured_articles

[4] There are also two classes for "list" type articles: Featured List and List-class. In this paper we do not examine list type articles.

[5] http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment#Grades

cles have the following suggestion: "Considerable editing is needed to close gaps in content and solve cleanup problems."

Complementing the quality assessment classes, Wikipedia has also developed a set of templates[6] for flagging articles having specific flaws, such as "this Wikipedia page is outdated" or "does not cite any references or sources". Unlike assessment classes, which provide explicit ratings of article quality with implicit suggestions for improvement, these templates explicitly express specific needed improvements that in turn implicitly describe a notion of article quality.

Our goal in this paper is to combine these two types of article quality labelling, by creating what we call an "actionable" quality model for the English Wikipedia. Such a model would both accurately assess article quality, and, through the features included in the model, explicitly indicate which aspects of the article need improvement.

1.2 Assessing quality

The research literature has both examples of studies looking at more general notions of quality (e.g., how collaboration leads to quality articles) and at specific features that make it possible to distinguish between high and low quality articles taken from the seven assessment classes (typically Featured Articles and Stub- or Start-class). We organise our review by whether the cited paper primarily describes editor- or article-based approaches, although some of them combine elements of both. We also discuss potential use cases that each approach might support.

1.2.1 Editor-based assessment

Editor-based approaches primarily investigate properties of the editors, for instance how much experience an editor has, in order to understand how those properties affect article quality.

Studies on *editor diversity* have looked at how the composition of editor experience, skills and knowledge, and coordination between editors determines article quality. For instance it has been shown that high quality articles have a large number of editors and edits, with intense cooperative behaviour [37]. Kittur and Kraut confirmed those results and refined them by discovering that article quality increases faster when a concentrated group of editors work together [20]. In that situation, explicit coordination by communicating on an article's talk page[7] is associated with a positive effect on quality. Arazy and Nov investigated further and found that editor concentration does not have a direct effect on article quality; instead it is indirect through editor coordination [3]. They also reported that article quality was directly affected by the article being edited by a diverse set of editors where at least some of them have a lot of Wikipedia experience. Lastly, Liu and Ram found that the different types of work editors do, as well as how they collaborate, affect article quality [22].

Editors' *reputation*, or history of making useful contributions, also affects article quality. Generally, research has calculated reputation or value metrics based on survival of article text on the word, paragraph, or revision level [1, 15, 32]. When looking at German Wikipedia, Stein and Hess found that editor reputation affected article quality from the first edit [27]. Similarly for English Wikipedia, Nemoto et al. found that articles started by editors with high social capital more quickly reached high quality status [24].

Research has also studied *article editor networks*, using the connections between editors and articles in order to understand how those connections affect quality. Brandes et al. and Wang and Iwaihara add several features describing editor activity to the edges in a network graph of an article and show that it is possible to predict article quality using statistics from this graph [5, 35]. A graph combining editors and articles was used by Wu et al., from which they mined certain network edge patterns between editors and articles, called *motifs*. They showed that the motifs could be used to predict article quality with comparable performance to other approaches [39].

These editor-based assessments help researchers understand how characteristics of articles and editors correspond to effective collaboration and could be used to study the overall quality of Wikipedia and how it has changed over time. They might also be directly useful for quality assessment work in Wikipedia, helping assessors find articles that are misclassified or that might be candidates for developing into Good or Featured Articles.

1.2.2 Article-based assessment

Editor-based metrics, though useful for modelling article quality, do not lead to directly actionable results. Knowing that editor diversity leads to higher article quality suggests that a user should look to collaborate with other editors with certain skills or experience levels, but does not explain how to create collaborations with those editors. Approaches that look at characteristics of articles (such as length) might be more useful for suggesting specific improvements that individual editors can make.

Wöhner and Peters investigated the life cycle of articles and found that high quality articles have an edit pattern that is distinctly different from low quality articles [38]. A shift to high quality happens in a burst, suggesting the shift requires a coordinated effort, as previously discussed when looking at editor diversity [20, 37].

Looking instead at article content, research has found that metrics describing the writing style of an article, for instance the variety of words used, correlates with quality [21, 40]. However, research has also shown that simply measuring the amount of content in the article by counting the number of words dominates when it comes to predictive power [4, 17, 38], suggesting that what Wikipedia articles mostly need to improve in quality is more content.

Some researchers use combinations of features describing the article's content as well as who edited it. Stvilia et al. mapped features from the information quality literature onto Wikipedia articles in a model combining aspects of authority/reputation and content that had good performance in distinguishing between Featured and Random articles [30]. Dalip et al. surveyed existing literature to collect a large number of features describing the article's content and other attributes of the article (e.g., average edits per day, or number of links from other articles) [16]. They found that the features describing the article's content (amount of content and writing style) were best able to distinguish between articles sampled from all assessment classes.

[6] `http://en.wikipedia.org/wiki/Wikipedia:Template_messages/Cleanup`

[7] A separate page for discussion of issues with the article.

1.3 Automatically detecting article flaws

As mentioned earlier, English Wikipedia has a number of cleanup templates that are used to describe specific flaws an article may have. These templates expose the flaws, making it easy to understand what tasks need to be done to improve the article. If we can automate detection of these flaws, we can use that to direct contributor attention. For instance, recommender systems that do Intelligent Task Routing [7] might show contributors articles like the ones they have already edited but which have specific flaws, and tell them which flaws need fixing.

Research on this subject is emerging and shows promise. Anderka et al. used binary classifiers for each flaw type and found good predictive performance for four of the ten types of flaws they investigated [2]. They were also involved with an international competition on quality flaw detection held in 2012, where an article corpus was gathered and two of the submitted classifiers showed promising performance [11].

One of the challenges facing this approach is that a single tag can indicate multiple flaws. For instance, to "wikify" an article might involve any of the ten tasks described by WikiProject Wikify[8]. This template was deprecated in January 2013 in favour of more specific templates[9], suggesting that Wikipedia contributors are also interested in having more clearly defined actionable tasks.

1.4 Our contribution

This paper contributes to the existing literature by combining quality assessment and flaw detection. Our approach has three goals:

1. **Specific Work Types:** A feature set that allows for identification of several types of work that can improve the quality of specific articles as well as fit the interests of a diverse population of contributors. For instance, some features may be well-suited for inexperienced editors to help with, while others might be a better fit for those with a lot of Wikipedia experience.

2. **Fine-Grained Classification:** It is useful to be able to distinguish between all seven assessment classes. Knowing the difference between complete (needs no additional work) and not really started (Stub-class) is not very useful. Finer-grained distinctions could be used to identify articles in need of reassessment to reduce lag in the assessment process, which currently relies solely on human intervention. We would also like to be able to describe the work that is needed to improve articles that already have a substantial amount of content.

3. **Efficient Analysis:** The model should not require extensive pre-analysis to work. Some existing models require analysis of the entire history of a Wikipedia, or build and mine complex data structures from an article's edit history. We prefer a solution that can run against readily available data and that is efficient enough to provide assessment and task recommendations for a large Wikipedia such as the English edition.

In the next section we describe how we developed our quality model, starting with Stvilia et al.'s model from 2005. We first add features that were less common when they developed that model, for instance citations of sources, and then prune non-actionable and poorly predicting features. We converge on a quality model with five actionable features—amount of content, number of citations, number of images, number of links to other Wikipedia articles, and article organisation with section headings—that classifies articles almost as well as models with many more, and non-actionable, features. Testing our model on all seven assessment classes also reveals that it usually estimates article quality to within one assessment class. We then discuss how this model and approach can be applied to Wikipedia and to what extent it might transfer to other language editions of Wikipedia, and perhaps other domains, where standards for quality can be different from the English edition.

2. AN ACTIONABLE QUALITY MODEL

In this section we will first discuss how we chose a machine learner that would enable us to learn more about which features were useful for predicting quality. We then discuss a simpler version of the assessment problem that divides the seven classes into a "good enough" set and one that "needs work", followed by how we created our training and test data sets using that definition. Using the chosen machine learner and data sets we then evaluate several different feature sets on our two-class problem, where we end up with our five-feature actionable quality model. We then test several different machine learners to understand if others significantly outperform our initial choice, before finally generalising our classification problem to all seven assessment classes.

2.1 Technology selection and data collection

2.1.1 Selecting an appropriate machine learner

Our aim is to gain an understanding of the predictive power of different features when classifying article quality, with particular focus on those that are actionable. We therefore prefer algorithms which allow us to inspect the underlying model directly. Blumenstock used a logistic regression where the regression coefficients are exposed [4]. Stvilia et al. used a decision tree classifier, where the tree can be inspected to learn how specific features are used, to build a fairly complex model with a combination of actionable and non-actionable features [30]. We chose to use a decision tree classifier because of the combination of an exposed model and known good performance from Stvilia et al.

2.1.2 Assessment class selection

A common approach to quality modelling is to classify Featured Articles (FAs) versus other articles (e.g., [30, 38]). However, we are interested in distinguishing broadly between articles that need a lot more attention and articles that are already "pretty good". Instead of predicting FAs versus others, we choose a split that reflects whether the articles are in need of more attention from contributors.

From the description of the assessment classes[10] we learn that both FAs and A-class articles are "complete". It is also clear that Good Articles (GA) have received a lot of attention, due to the peer review process involved in reaching

[8] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Wikify

[9] See http://en.wikipedia.org/wiki/Template:Wikify

[10] http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment#Grades

GA status. Thus, we choose to split the article space into two classes: one class of articles not in need of more attention, which we label *GoodEnough*, containing FA, GA and A-class articles, and one class of articles needing more attention, which we label *NeedsWork*, containing B-, C-, Start-, and Stub-class articles. Because we include all classes of articles, and set our split not at the best (FA) or worst (Stub and Start) article classes, but somewhere in the middle, we expect this to be a challenging task.

2.1.3 Data collection

Having chosen a decision tree classifier as our technology and defined our classes as "GoodEnough = FA + GA + A" and "NeedsWork = B + C + Start + Stub", we turn our attention to gathering articles for training and testing the classifier. Decision tree classifiers prefer training sets where there are roughly the same number of items in each class, so we set out to build such a dataset.

We gathered our data in the period of 27-29 May 2011. First we found the class with the fewest number of articles, which was A-class articles with 827[11]. Our plan was to select the same number of articles from FA, GA, and A, leading to a total of 2481 articles in the *GoodEnough* class, then sample another 2481 evenly distributed across the B, C, Start, and Stub classes to create a balanced dataset of roughly 5000 articles. However, when we crawled A-class articles using the category "Category:A-class articles", we found only 672 actual articles[12].

In the end, we gathered the 672 A-class articles and 800 each from FAs and GAs for a total of 2272 *GoodEnough* articles. We then chose 568 articles from each of the remaining four classes, for a total of 2272 *NeedsWork* articles. These were then split 50/50 into a training set and a test set. Note that these assessments are best guesses; a limitation of this data set is that the quality assessment assigned to articles may not reflect their true assessment class, or the underlying distributions in Wikipedia, because articles change in quality and some articles are not assessed.

2.2 Establishing a baseline

We start our exploration of article quality assessment with Stvilia et al.'s early but well-known model as a baseline. This model was chosen because it has known good performance and contains a combination of actionable and less actionable features. There are a total of 18 features in all, some of which are added together to make it a seven-dimensional model as presented below.

1. **Authority/Reputation** = 0.2*NumUniqueEditors + 0.2*NumEdits + 0.1*Connectivity + 0.3*NumReverts + 0.2*NumExternalLinks + 0.1*NumRegUserEdits + 0.2*NumAnonEdits
2. **Completeness** = 0.4*NumBrokenWikilinks + 0.4*NumWikilinks + 0.2*ArticleLength
3. **Complexity** = Flesch-Kincaid Readability Score
4. **Informativeness** = 0.6*InfoNoise - 0.6*Diversity + 0.3*NumImages
5. **Consistency** = 0.6*AdminEditShare + 0.5*Age
6. **Currency** = Current article age in days
7. **Volatility** = Median revert time in minutes

Connectivity is the number of articles reachable through the editors of a given article. *InfoNoise* is the proportion of text content remaining after removing MediaWiki code and stopwords and stemming all words. *Diversity* is *NumUniqueEditors/NumEdits*. Other definitions can be found in the original paper [30].

Note that some features, such as the current article age or revert volatility, are practically impossible to directly change; others, involving the mix of anonymous-to-registered or admin-to-regular edits, are in principle actionable by recruiting new editors (or suppressing current ones) but in practise difficult for individuals to enact; while still others, such as the number of wikilinks[13] or images, might be more directly addressable by individual editors.

In order to identify reverts to calculate *Volatility*, we applied the approach of Priedhorsky et al., which uses regular expressions to match edit comments [25]. Edit comments are a text field used by contributors to describe the changes they have made in a revision. While this approach does not correctly identify all reverts [10], in May 2011 when we collected our dataset more resilient approaches would have required downloading the text of all revisions of each article to calculate hash values[14]. We also used Priedhorsky et al.'s approach to identify anti-vandal work and exclude anti-vandal edits from median revert time, as much vandal fighting is now handled by bots and software-assisted humans [13] and therefore does not properly reflect article controversy. Bot edits were identified by making a case-insensitive match of the username associated with the edit having a part that ends with "bot", for example "RamBot" and "MiszaBot III". The advantage of this approach is that it is fast, but it will miss bots that do not follow the common naming convention of bot accounts[15]. Checking if the account is a member of the "bot" user group should catch most or perhaps all of the missed bots that are officially registered with Wikipedia.

The *Connectivity* feature is the cardinality of the set of other articles edited by the editors of a specific article, after excluding bots and anti-vandal reverting editors from the set. At the time we gathered our data it was nontrivial to determine how reverts affect an article's history [9, 10]; thus we did not attempt to remove reverted editors when looking for connected articles.

Some of the data used in Stvilia et al.'s model is power law distributed, e.g., number of edits and number of editors. The paper did not specify whether they chose to log-transform these features, so we tested the model with both non-transformed and log-transformed. Non-transformed data had higher performance so we report it here.

We test this set of features using the C4.5 decision tree classifier and our training and test datasets described earlier. The overall classification results are listed in the "2005 model" column in Table 1. We report the following measures: True Positive Rate (**TPR**) for each class as well as an

[11] This was according to "WP 1.0 Bot", which counts the number of articles in each quality class.

[12] The difference appears to come from "WP 1.0 Bot" using WikiProject listings of article assessment instead of counting articles tagged with this category.

[13] Links to other Wikipedia articles.

[14] SHA1 hash values for all revisions are now available through Wikipedia's API.

[15] `http://en.wikipedia.org/wiki/WP:BOTACC`

overall weighted average, which allow us to judge the classifier's ability to correctly predict classes; **Precision** and **Recall**, which are widely used to judge performance when one class is more important; **F-Measure**, which represents a harmonic mean between precision and recall; and **ROC** (Receiver Operating Characteristic), which is commonly used to judge relative performance between classifiers for the trade-off between true positive and false positive rates.

Because we defined our "GoodEnough" class to include Featured Articles, Good Articles, and A-class articles, while Stvilia et al. classified Featured Articles versus Random with Stub-class articles removed, we expect to see somewhat lower performance compared to theirs. As we see from Table 1, overall prediction performance comes in at 76.1%, while in their work they successfully classified over 90% of their articles.

2.3 New potential features

In addition to the features used in Stvilia et al.'s model, we are interested in introducing new features, including actionable features that suggest specific improvements and features that have become more common in Wikipedia since 2005. For instance we know that Wikipedia articles require sources for claims[16]. Previous research has shown that when readers judge the trustworthiness of Wikipedia articles, references to sources play an important part [23]. To capture the extent to which claims in the article are sourced we propose *NumReferences*, a measure of the number of citations, by counting the number of <ref>-tags which are used for footnote citations.

We also add a feature to capture the extent to which an article has been organised into sections (*NumHeadings*). Appropriate article structure and organisation is a common theme in the article assessment criteria and many Wikipedia articles have sections such as "See also" for linking to other relevant Wikipedia articles and "References" for listing the article's sources. Research has suggested that organising content in a wiki can help structure future contributions [26], meaning this feature can both reflect current article quality and improve future contributions.

Some of these added features might be good metrics by themselves, but it could also be that there is a relationship to the length of the article. For instance the raw number of cited claims is likely to be lower for a short article, but it might be that relative to its length it has an appropriate number of citations. We therefore add features to capture the relationship with article length, as in de la Calzada and Dekhtyar [8]: *NumReferences/ArticleLength*, *NumImages/ArticleLength*, *NumWikilinks/ArticleLength*, and *NumHeadings/ArticleLength*.

Because many good articles have an infobox, we add a binary (0/1) categorical feature for that. Lastly we add features for the number of templates and categories an article has (*NumTemplates* and *NumCategories*). High quality articles are likely to use templates for formatting of content and following Wikipedia conventions, whereas low quality articles might lack these. Similarly we suspect that high quality articles will be assigned to a number of categories, whereas low quality articles may be less likely to be categorised well.

We also propose an editor tenure metric to replace Stvilia et al.'s administrator edit share because the proportion of administrators to other contributors on English Wikipedia is now much lower [31]. While this is not an actionable feature, we are interested in understanding its effect on performance as previous research suggests that edits by experienced editors positively affect article quality [24, 27]. We want to capture a notion of total editor experience accumulated across all edits to an article, in both age (time since they registered) and number of edits. This leads us to log-transform the edit count, because it is known to be power-law distributed, and then linearly combine them for each edit a user makes to a specific article as follows:

$$tenuretime(t,i) = t - t_{reg,i} \quad (1)$$

$$tenureedits(t,i) = log(nedits_{i,now} * t/(t_{now} - t_{reg,i})) \quad (2)$$

$$tenure(t,i) = tenuretime(t,i) + tenureedits(t,i) \quad (3)$$

In the formulae above, t is the time user i edited the article, $t_{reg,i}$ is the time user i registered their account, while $nedits_{i,now}$ is user i's edit count as of when the calculation was done (t_{now}). We then sum $tenure(t,i)$ for all registered non-reverting, non-bot editors of a given article to get our proposed metric *Tenure*.

2.4 Building new models

We now turn our attention to investigating how different feature sets and machine learning technologies affect classification performance. We first modify two of Stvilia et al.'s features and add our proposed ones to create and evaluate a large model with 17 features. Then we describe how we iteratively tested and removed specific features to create a hybrid model with eight features, ending up with a model that only contains five actionable features. Lastly we evaluate the performance of other classifiers.

2.4.1 Full model

We start by modifying the seven dimensions so that the features become more clearly separated between the actionable and non-actionable, then add our proposed features. Separating *ArticleLength* from *Completeness* leaves it a measure of the number of wikilinks, and removing *Diversity* from *Informativeness* leaves that feature a measure of textual noise and number of images. The resulting model contains 17 dimensions, as previously defined unless noted, and will be referred to as the "full model". Table 2 lists all features ranked by their overall gain ratio as calculated by WEKA using cross-validation on the training set. Gain ratio is the measure used in a C4.5 decision tree to determine which feature to use when splitting between classes [33].

Training WEKA's C4.5 decision tree classifier using these 17 features results in a tree of size 153 with 77 leaves. It correctly classifies 1951 articles, or 85.9%, as shown in the "Full model" column in Table 1. This large increase in performance comes mainly from the *NeedsWork* class, which the seven feature model only correctly classified 68.3% of the time, while the full model correctly predicted 86.8% of the articles in that class.

2.4.2 Hybrid model

One of the reasons for choosing to use a decision tree was the ability to inspect the tree to understand how the features were used and whether some would be good candidates for removal. Inspecting the tree trained on the full

[16] `http://en.wikipedia.org/wiki/Wikipedia:Verifiability`

	2005 model	Full model	Hybrid model	Actionable model
Features	Authority/Reputation	Authority/Reputation	Authority/Reputation	
	Completeness	Completeness*	Completeness*	Completeness*
	Complexity	Complexity		
	Informativeness	Informativeness*	Informativeness*	Informativeness*
	Consistency	Consistency		
	Currency	Currency		
	Volatility	Volatility		
		ArticleLength	ArticleLength	ArticleLength
		Diversity	Diversity	
		Tenure	Tenure	
		NumHeadings	NumHeadings	NumHeadings
		NumRefs/Length	NumRefs/Length	NumRefs/Length
		NumReferences		
		NumHeadings/Length		
		NumImages/Length		
		NumWikilinks/Length		
		HasInfobox		
***GoodEnough* TPR**	0.839	0.849	0.899	0.898
***NeedsWork* TPR**	0.683	0.868	0.854	0.833
Overall TPR	0.761	0.859	0.876	0.865
Precision	0.767	0.859	0.877	0.867
Recall	0.761	0.859	0.876	0.865
F-measure	0.760	0.859	0.876	0.865
ROC	0.792	0.863	0.884	0.883

Table 1: Feature list and overall classification results of all four models. Features marked * are modified as described in section 2.4.3. For a description of classification performance measures, see section 2.2.

model, we found that one feature was never used (*NumWikilinks/ArticleLength*) while some features (e.g., *Authority*, *Complexity*, and *Currency*) were mainly used in deep branches to distinguish between a small number of articles. We saw these features as likely candidates for removal to prevent over-fitting without a large impact on performance.

We also iteratively added and evaluated specific features or combinations of these as an alternative to a large feature set that leaves the classifier to figure out which ones are useful. The complete process is omitted for brevity, consisting of testing more than 30 models with various combinations of features. We kept features that created fairly simple trees, indicating they had good information gain, while performing on par with classification performance using the full feature set. *Complexity*, *Volatility*, and *Currency* were removed without impacting performance. The *Consistency* feature was dropped in favour of *Tenure*. The result is our "hybrid model" with eight features, combining actionable and non-actionable ones: *Authority/Reputation*, *Completeness*, *Informativeness*, *Diversity*, *Tenure*, *ArticleLength*, *NumHeadings*, and *NumReferences/ArticleLength*.

The "Hybrid model" column in Table 1 shows the overall performance of this hybrid model being slightly better than the one trained on the full list of features. It is correctly identifying more *GoodEnough* articles (89.9% compared to 84.9%) at the cost of misclassifying some additional *NeedsWork* articles (14.6% compared to 13.2%).

2.4.3 Actionable model

Because of our interest in actionable features, we next looked at the impact of removing all remaining non-actionable features from the model, resulting in our "actionable model" which contains only five dimensions[17]:

1. Completeness = 0.4*NumBrokenWikilinks + 0.4*NumWikilinks
2. Informativeness = 0.6*InfoNoise + 0.3*NumImages
3. NumHeadings
4. ArticleLength
5. NumReferences/ArticleLength

The "Actionable model" column in Table 1 shows that this model has comparable performance to the full and hybrid models. Our actionable model incorrectly regards a slightly larger proportion of *NeedsWork* articles as high quality. This could be due to a lag in the assessment process, as discussed earlier: it may be that articles edited by high-profile editors are more likely to be reassessed. It could also be that those articles are more likely to be of high quality, as we argued when defining our *Tenure* metric. Since our five-feature model does not contain features for editor experience, it will instead regard articles as high quality based purely on content features.

2.5 Alternative classifiers

The decision tree was useful for exploring and selecting features, and though it provided good performance, other classifiers might outperform it. We used some of WEKA's other available classifiers, including libSVM (Support Vector Machine), MultilayerPerceptron (neural network), JRip

[17] For the definition of *InfoNoise*, see section 2.2.

Rank	Feature	Overall gain ratio	Actionable
1	NumReferences/ArticleLength	0.205 ± 0.018	Yes
2	NumReferences	0.190 ± 0.012	Yes
3	ArticleLength	0.159 ± 0.015	Yes
4	Diversity	0.135 ± 0.006	No
5	Tenure	0.123 ± 0.005	No
6	NumHeadings	0.114 ± 0.007	Yes
7	NumHeadings/ArticleLength	0.105 ± 0.004	Yes
8	Informativeness = 0.6*InfoNoise + 0.3*NumImages	0.101 ± 0.005	Yes
9	Completeness = 0.4*NumBrokenWikilinks + 0.4*NumWikilinks	0.101 ± 0.003	Yes
10	NumImages/ArticleLength	0.099 ± 0.002	Yes
11	NumWikilinks/ArticleLength	0.091 ± 0.003	Yes
12	Authority/Reputation	0.081 ± 0.003	No
13	Consistency	0.055 ± 0.002	No
14	Volatility	0.043 ± 0.002	No
15	Currency	0.025 ± 0.002	No
16	HasInfobox	0.018 ± 0.002	Yes
17	Complexity	0.016 ± 0.002	Yes

Table 2: Overall gain ratio evaluation for all 17 features.

Classifier	*GE* TPR	*NW* TPR	Overall TPR	Precision	Recall	F-measure	ROC
RandomForest	0.889	0.856	0.872	0.873	0.872	0.872	0.939
C4.5	0.898	0.833	0.865	0.867	0.865	0.865	0.883
MultiLayerPerceptron	0.889	0.824	0.857	0.858	0.857	0.856	0.904
JRip	0.882	0.800	0.841	0.843	0.841	0.841	0.871
LibSVM	0.886	0.662	0.774	0.789	0.774	0.771	0.774
SimpleLogistic	0.824	0.708	0.766	0.769	0.766	0.765	0.843

Table 3: Classification results for all classifiers, actionable model with five features, ranked by F-measure. *GE* TPR and *NW* TPR are True Positive Rate for the *GoodEnough* and *NeedsWork* class, respectively.

(rule-based), SimpleLogistic (logistic regression), and RandomForest with 100 trees. All classifiers used WEKA's default options, with the exception of the random forest, which was tested with sizes from 10 (the default) up to 1000. We report results based on a random forest size of 100 as it had the best performance.

We tested both the full model with 17 features and the 5-feature actionable model[18]. The results for both feature sets were comparable, with only minor improvements in both cases, so we report results for the actionable model in Table 3. These results indicate that we might need a different set of features to tease out the benefit of specific classifiers, something which future research could look into.

2.6 Predicting all assessment classes

Our investigation of actionable features is motivated by our interest in using those features to help contributors increase the quality of articles. Being able to distinguish between all seven assessment classes could support other quality-related use cases. We might be able to identify articles that need reassessment (e.g., candidates to become Featured Article), allow users to focus on particular quality levels (e.g., avoiding Stub-class articles or looking for articles near the borderline of quality classes), or highlight ways the classes differ on specific features. Distinguishing between all seven classes has also received relatively little research attention, despite its interestingness as a problem. While the difference between a Featured Article and a Stub-class article may be large enough to make it straightforward to differentiate between them, the boundaries between some of the other classes (e.g., between C-class and B-class) are likely to be less well-defined because there are smaller differences in the assessment criteria and because of errors and lag in assessment.

In these evaluations we reuse our existing training and test datasets, but do not collapse them into two classes. We again evaluate both the full model with 17 features and the actionable model with five features; as before, the results are comparable. We also tested all of the classifiers described in the last section, and again the random forest classifier was the highest-performing classifier. Thus, below we report only on the results for the random forest classifier using the five-feature model. We then discuss how the results differ depending on the classifier and feature set.

Table 4 shows the performance of a random forest classifier with 100 trees using the actionable model with five features to classify all seven classes. In this table we also report the false positive rate (**FPR**), which is the proportion of other articles predicted to belong to a given class and allows us to judge the confusion between classes. Overall the classifier only correctly classifies 42.5% of the articles, showing that this is a very difficult classification problem. Some of the classes are easier to predict than others, with performance on Featured Article (FA) and Stub-class of 60.3% and 57.7%, respectively. As we speculated above, results are worst in the middle for A-, B-, and C-class articles.

The full confusion matrix is shown in Table 5. Two important patterns emerge from this matrix. First is that there

[18] The SVM classifier was not run on the large feature set as the high dimensionality leads to poor performance.

Class	**TPR**	**FPR**	**Precision**	**Recall**	**F-measure**	**ROC**
FA	0.603	0.165	0.439	0.603	0.508	0.857
GA	0.433	0.126	0.424	0.433	0.428	0.806
A	0.289	0.079	0.388	0.289	0.331	0.733
B	0.327	0.096	0.327	0.327	0.327	0.764
C	0.292	0.102	0.290	0.292	0.291	0.772
Start	0.405	0.088	0.398	0.405	0.401	0.825
Stub	0.577	0.021	0.796	0.577	0.669	0.934
Overall	0.425	0.101	0.436	0.425	0.425	0.813

Table 4: Classification results per class for all seven assessment classes, using the actionable model with five features and a random forest classifier with 100 trees.

	FA	*GA*	*A*	*B*	*C*	*Start*	*Stub*	*No. articles*
FA	241	83	62	11	2	1	0	400
GA	128	173	58	21	17	3	0	400
A	123	63	97	19	18	12	4	336
B	29	40	18	93	74	27	3	284
C	22	37	9	72	83	51	10	284
Start	6	11	6	50	71	115	25	284
Stub	0	1	0	18	21	80	164	284
No. articles	549	408	250	284	286	289	206	2272

Table 5: Confusion matrix for classification of all seven assessment classes, actionable model with five features, random forest classifier with 100 trees. Bold rows show correct class, italic columns show predicted class. The highlighted diagonal shows correctly classified articles.

is a lot of confusion between FA, GA, and A-class articles. Both FA and A-class articles are defined as "complete", thus they should mostly differ by what comes out of the FA review process. Our model does not appear to capture that difference, with 123 of the 400 A-class articles (30.1%) predicted to be Featured Articles.

The second pattern is that the classifier is pretty good at getting within one class, and tends to err on the high side. If we allow the classifier to be off by one class[19], it correctly identifies 1747 articles, or 76.9%. This still might be useful for human-in-the-loop tasks such as reviewing quality assessments, but probably does not perform well enough for automatic tasks such as filtering articles out of suggestion lists based on quality class.

If we relax the requirement that features be actionable, and test the full 17-feature model, we see a gain in overall performance from 42.5% to 48.3%. FA and GA are the classes with large gains, improving their true positive rate to 74% and 57%, respectively. The other classes see little or no improvement, indicating that distinguishing between the remaining classes requires other types of features. This suggests that use cases which do not need actionability, such as assessment, would benefit from using a richer feature set.

When it comes to performance differences between different types of classifiers, we found that some perform very well on certain classes. The neural network and rule-based classifiers performed well on the Featured Article class, but the latter struggles with C-class articles. We see this as a good opportunity for future research to look at ensemble methods to exploit the advantages of some of the classifiers when it comes to predicting specific assessment classes.

This completes our model-building exploration. We have found that a simple set of five numeric features[20] provides good performance for assessing Wikipedia article quality using a decision tree or random forest classifier. Our initial problem divided English Wikipedia's seven assessment classes into two classes depending on whether the articles appeared to need more attention or not, but we also saw promising performance on predicting all seven classes.

3. DISCUSSION

As we saw in the previous section, building an actionable quality model for Wikipedia is realisable. We now discuss the impact and limitations of our findings, starting with revisiting the three goals we defined in the introduction.

3.1 Meeting model goals

Our first goal, *Specific Work Types*, sought "a feature set that allows for identification of several types of actionable work." We ended up with a model that contains five features, of which two relate to the overall content and its organisation (*ArticleLength* and *NumHeadings*), while the three others relate to specific content features, including wikilinks, images, and citations. The model should therefore cater to both those who are interested in researching content to add to an article, as well as those interested in more precisely defined tasks such as adding an image or finding sources for specific claims.

Goal number two, labelled *Fine-Grained Classification*, was "the ability to evaluate between quality of articles across all seven assessment classes." We found it feasible to predict multiple classes, although with a fair amount of uncertainty. As we will discuss below, this problem has not received much attention in the research literature and is an area where we see several opportunities for valuable contributions.

Our last goal, *Efficient Analysis*, required that our model "does not require system-wide pre-analysis to work." Unlike some other techniques which require computation or maintenance of large amounts of historical data, or which need

[19] Where FA=FA or A; A=FA, A, or B; Stub=Stub or Start.

[20] The specific features are listed at the end of section 2.4.

data about an article's edit history, our model only requires readily available descriptive measures of an article's current text and does not perform complex analysis. Because it does require access to the article text, performance will be limited by how quickly that can be retrieved. Our current implementation is written in Python with a Java-based XML-RPC service for classification, and uses Wikipedia's API for text retrieval. It uses the Wikimedia Toolserver[21] for wikilink and image metadata, and processes about three articles per second. It is therefore currently able to support non-realtime tasks on the scale of assessing quality for a WikiProject with thousands of articles. While it can easily be parallelised for further speed improvements, it might still not be fast enough for real-time assessment on the revision level of large Wikipedias.

3.2 Uses, improvements, and limitations

We now turn our attention to possible uses of our findings, their limitations and implications for both the English edition and Wikipedia in general, and potential directions for future research.

An actionable quality model can be turned on its head; instead of being used to classify articles it can calculate how a contribution affects quality and thereby measure the value of contributions. This can be used to estimate the value of a single revision, which can then be aggregated to calculate value on the article, editor, WikiProject, or community level. Existing measures of contributor value include counting the number of edits [19], variations on survival of contributions [1, 15, 25], and estimated amount of labour hours [12]. Our approach is more directly tied to article quality than any of these. Future research could look into its suitability for measuring editor or contribution value and how exposing these notions of value affect the community, or whether it allows for catching borderline cases of vandalism not already caught by vandal-fighting tools.

As we have seen, our model allows for predicting quality using all seven assessment classes, although with a fair amount of uncertainty. What is the effect of showing assessment of an article's quality on people's willingness to contribute? Keegan and Gergle found that quality begets edits (i.e., higher quality leads to more edits) [18]. Does this also occur if the contributors are aware of the assessed quality before they decide to edit? Their willingness to contribute might also be related to how many views the article gets; we have been running live experiments on English Wikipedia aiming to find answers to these questions.

In addition to overall quality assessment, we can use the model to show which tasks are needed and how much each of them might improve quality. Exposing this information might also motivate contributors to edit. Task information could be aggregated into a pool of articles needing specific tasks, allowing us to make recommendations based on work type rather than, or in addition to, article topic, which might increase the utility of recommendations from tools such as SuggestBot [7]. Previous research has shown that some Wikipedia editors prefer to do particular kinds of tasks some of the time [36], which work type-based suggestions could be a good fit for. More generally, predicting the kinds of work an editor likes to do would be an interesting research problem that actionable quality models might help address.

While our actionable quality model has a healthy blend of features, a potential drawback is that none of the tasks associated with these features may be well suited for inexperienced Wikipedia editors. Newcomers might instead prefer tasks that are easier, shorter, or require less Wikipedia-specific knowledge, such as copy edits. Our model did start out with the Flesch-Kincaid readability index as one feature, but as in Dalip et al. [16], readability was found to not be a strong predictor. Future work could investigate what tasks are performed by users depending on their experience level and to what extent it is possible to detect if an article has flaws matching those tasks.

We investigated the feasibility of predicting article quality across all of English Wikipedia's seven assessment classes. Our results indicate that doing so accurately is difficult and that there are likely subtle differences between some classes, but we found that prediction to within one assessment class is feasible. Reviewing the existing literature suggests that this problem has not seen extensive interest, even though these classes have been in use for several years. This can therefore be a promising venue for future research, perhaps through gathering a vetted data set and holding a classification contest similar to the Wikipedia participation challenge in 2011[22] and the flaw detection contest held in 2012 [11].

Doing this could provide a better understanding of quality differences between assessment classes. Certain features may be most predictive for certain classes, or particular kinds of work may be associated with moving between specific assessment levels. Improving an article from Stub- to Start-class might require vastly different actions than moving from A-class to Featured Article. Are the features at higher levels more subtle than the ones we discovered, where "add more content" and "add sources" dominated? These investigations might also lead to methods that allow us to measure the evolution of Wikipedia quality over time.

The ability to predict all seven assessment classes could also allow us to identify articles that appear to be incorrectly assessed, which we could then expose to contributors for consideration. As described in section 2.1.2, a limitation of our data set is how accurate the assessments are. For instance it could be that different users have different mental models of what a B-class article is, or that an article has substantially changed since it was assessed. A tool that allows easy access to likely candidates for reassessment could potentially both reduce assessment lag while also making articles within each assessment class more homogeneous.

Finally, looking at how different language editions approach the problem quality might be fruitful. As discussed in the introduction, the notion of quality differs between languages [28], something we experienced firsthand when we attempted to apply our classifier to the Swedish and Norwegian language editions. Where English Wikipedia uses footnote citations extensively, these languages instead often have a bibliography section and no inline citations, rendering our *NumReferences/ArticleLength* feature useless. While adopting the English Wikipedia's use of footnotes for citations could unify behaviour and allow our model to work across languages, each language also reflects cultural heritage for encyclopaedias, making it worth discussing to what extent such unification would be desirable.

[21] http://toolserver.org

[22] https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2011-10-31/Recent_research

4. CONCLUSION

Our initial goal was to combine article quality assessment and flaw detection. As we have seen, this was successful, resulting in a simple model of article quality with actionable features. Making this technology available to the Wikipedia community can enable easier access to assessment, as well as suggesting specific tasks for improvement, and through that help improve the quality of Wikipedia articles.

5. ACKNOWLEDGEMENTS

We would like to thank the other members of GroupLens Research for their support, the anonymous reviewers for their helpful suggestions, the Wikimedia Foundation and Wikimedia Deutschland for facilitating access to Wikipedia data, and all Wikipedia contributors for creating this great encyclopaedia. This work is funded by the National Science Foundation (grants IIS 08-08692, 09-68483, 08-45351).

6. REFERENCES

[1] B. T. Adler and L. De Alfaro. A content-driven reputation system for the Wikipedia. In *Proc. WWW*, 2007.

[2] M. Anderka, B. Stein, and N. Lipka. Predicting quality flaws in user-generated content: the case of Wikipedia. In *Proc. SIGIR*, pages 981–990, 2012.

[3] O. Arazy and O. Nov. Determinants of Wikipedia quality: the roles of global and local contribution inequality. In *Proc. CSCW*, pages 233–236, 2010.

[4] J. E. Blumenstock. Size matters: word count as a measure of quality on Wikipedia. In *Proc. WWW*, 2008.

[5] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in Wikipedia. In *Proc. WWW*, pages 731–740, 2009.

[6] T. Chesney. An empirical examination of Wikipedia's credibility. *First Monday*, 11(11):1–13, 2006.

[7] D. Cosley, D. Frankowski, L. Terveen, and J. Riedl. SuggestBot: using intelligent task routing to help people find work in Wikipedia. In *Proc. IUI*, pages 32–41, 2007.

[8] G. De la Calzada and A. Dekhtyar. On measuring the quality of Wikipedia articles. In *Proc. WICOW*, pages 11–18, 2010.

[9] M. D. Ekstrand and J. T. Riedl. rv you're dumb: identifying discarded work in Wiki article history. In *Proc. WikiSym*, pages 4:1–4:10, 2009.

[10] F. Flöck, D. Vrandečić, and E. Simperl. Revisiting reverts: accurate revert detection in Wikipedia. In *Proc. HT*, pages 3–12, 2012.

[11] P. Forner, J. Karlgren, and C. E. Womser-Hacker. Overview of the 1st International Competition on Quality Flaw Prediction in Wikipedia. *CLEF 2012 Evaluation Labs and Workshop*, 2012.

[12] R. S. Geiger and A. Halfaker. Using edit sessions to measure participation in Wikipedia. In *Proc. CSCW*, pages 861–870, 2013.

[13] R. S. Geiger and D. Ribes. The work of sustaining order in Wikipedia: the banning of a vandal. In *Proc. CSCW*, pages 117–126, 2010.

[14] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[15] A. Halfaker, A. Kittur, R. Kraut, and J. Riedl. A jury of your peers: quality, experience and ownership in Wikipedia. In *Proc. WikiSym*, pages 15:1–15:10, 2009.

[16] D. Hasan Dalip, M. André Gonçalves, M. Cristo, and P. Calado. Automatic quality assessment of content created collaboratively by web communities: a case study of Wikipedia. In *Proc. JCDL*, pages 295–304, 2009.

[17] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. Measuring article quality in Wikipedia: models and evaluation. In *Proc. CIKM*, pages 243–252, 2007.

[18] B. Keegan and D. Gergle. Egalitarians at the gate: one-sided gatekeeping practices in social media. In *Proc. CSCW*, pages 131–134, 2010.

[19] A. Kittur, E. Chi, B. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *alt.CHI at CHI*, 2007.

[20] A. Kittur and R. E. Kraut. Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In *Proc. CSCW*, pages 37–46, 2008.

[21] N. Lipka and B. Stein. Identifying featured articles in Wikipedia: writing style matters. In *Proc. WWW*, pages 1147–1148, 2010.

[22] J. Liu and S. Ram. Who does what: Collaboration patterns in the Wikipedia and their impact on article quality. *ACM TMIS*, 2(2):11:1–11:23, July 2011.

[23] T. Lucassen and J. M. Schraagen. Trust in Wikipedia: how users trust information from an unknown source. In *Proc. WICOW*, pages 19–26, 2010.

[24] K. Nemoto, P. Gloor, and R. Laubacher. Social capital increases efficiency of collaboration among Wikipedia editors. In *Proc. HT*, pages 231–240, 2011.

[25] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in Wikipedia. In *Proc. GROUP*, pages 259–268, 2007.

[26] J. Solomon and R. Wash. Bootstrapping wikis: developing critical mass in a fledgling community by seeding content. In *Proc. CSCW*, pages 261–264, 2012.

[27] K. Stein and C. Hess. Does it matter who contributes: A study on featured articles in the German Wikipedia. In *Proc. HT*, pages 171–174, 2007.

[28] B. Stvilia, A. Al-Faraj, and Y. J. Yi. Issues of cross-contextual information quality evaluation–The case of Arabic, English, and Korean Wikipedias. *Library & Information Science Research*, 31(4):232–239, 2009.

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

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

[31] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Information quality work organization in Wikipedia. *Journal of ASIST*, 59(6):983–1001, 2008.

[32] Y. Suzuki and M. Yoshikawa. Mutual evaluation of editors and texts for assessing quality of Wikipedia articles. In *Proc. WikiSym*, 2012.

[33] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson Education, 2006.

[34] F. B. Viégas, M. Wattenberg, and M. M. McKeon. The hidden order of Wikipedia. In *Online communities and social computing*, pages 445–454. 2007.

[35] S. Wang and M. Iwaihara. Quality evaluation of Wikipedia articles through edit history and editor groups. In *Web Technologies and Applications*, pages 188–199. 2011.

[36] M. Wattenberg, F. B. Viégas, and K. Hollenbach. Visualizing activity on Wikipedia with chromograms. In *Human-Computer Interaction–INTERACT 2007*, pages 272–287. Springer, 2007.

[37] D. M. Wilkinson and B. A. Huberman. Cooperation and quality in Wikipedia. In *Proc. WikiSym*, 2007.

[38] T. Wöhner and R. Peters. Assessing the quality of Wikipedia articles with lifecycle based metrics. In *Proc. WikiSym*, pages 16:1–16:10, 2009.

[39] G. Wu, M. Harrigan, and P. Cunningham. Classifying Wikipedia articles using network motif counts and ratios. In *Proc. WikiSym*, 2012.

[40] Y. Xu and T. Luo. Measuring article quality in Wikipedia: Lexical clue model. In *3rd Symposium on Web Society*, pages 141–146. IEEE, 2011.