Learning to Predict the Quality of Contributions to Wikipedia

Gregory Druck and **Gerome Miklau** and **Andrew McCallum**
{gdruck,miklau,mccallum}@cs.umass.edu
Department of Computer Science
University of Massachusetts
Amherst, MA 01003

Abstract

Although some have argued that Wikipedia's open edit policy is one of the primary reasons for its success, it also raises concerns about quality — vandalism, bias, and errors can be problems. Despite these challenges, Wikipedia articles are often (perhaps surprisingly) of high quality, which many attribute to both the dedicated Wikipedia community and "good Samaritan" users. As Wikipedia continues to grow, however, it becomes more difficult for these users to keep up with the increasing number of articles and edits. This motivates the development of tools to assist users in creating and maintaining quality. In this paper, we propose metrics that quantify the quality of contributions to Wikipedia through implicit feedback from the community. We then learn discriminative probabilistic models that predict the quality of a new edit using features of the changes made, the author of the edit, and the article being edited. Through estimating parameters for these models, we also gain an understanding of factors that influence quality. We advocate using edit quality predictions and information gleaned from model analysis not to place restrictions on editing, but to instead alert users to potential quality problems, and to facilitate the development of additional incentives for contributors. We evaluate the edit quality prediction models on the Spanish Wikipedia. Experiments demonstrate that the models perform better when given access to content-based features of the edit, rather than only features of contributing user. This suggests that a user-based solution to the Wikipedia quality problem may not be sufficient.

Introduction and Motivation

Collaborative content generation systems such as Wikipedia are promising because they facilitate the integration of information from many disparate sources. Wikipedia is remarkable because anyone can edit an article. Some argue that this open edit policy is one of the key reasons for its success (Roth 2007; Riehle 2006). However, this openness does raise concerns about quality — vandalism, bias, and errors can be problems (Denning et al. 2005; Riehle 2006; Kittur et al. 2007).

Despite the challenges associated with an open edit policy, Wikipedia articles are often of high quality (Giles 2005). Many suggest that this is a result of dedicated users that make many edits, monitor articles for changes, and engage in debates on article discussion pages. These users are sometimes referred to as "zealots" (Anthony, Smith, and Williamson 2007), and studies claim that they are motivated by a system of peer recognition that bears resemblance to the academic community (Forte and Bruckman 2005). However, the contributions of "good Samaritan" users, who edit articles but have no desire to participate in the community, cannot not be underestimated (Anthony, Smith, and Williamson 2007).

As Wikipedia continues to grow, however, it becomes more difficult for these users to keep up with the increasing number of articles and edits. Zealots comprise a relatively small portion of all Wikipedia users. Good Samaritan users are not likely to seek out errors, but instead rely on stumbling upon them. It is interesting to consider whether aiding users in detecting and focusing effort on quality problems could improve Wikipedia.

In this paper, we examine the problem of estimating the quality of a new edit. Immediately, we face the problem of defining edit quality. It has been argued that there is no general definition of information quality, and hence quality must be defined using empirical observations of community interactions (Stvilia et al. 2008). Therefore, we define quality using implicit feedback from the Wikipedia community itself. That is, by observing the community's response to a particular edit, we can estimate the edit's quality. The quality metrics we propose are based on the assumption that edits to an article that are retained in subsequent versions of the article are of high quality, whereas edits that are quickly removed are of low quality.

We use these community-defined measures of edit quality to learn statistical models that can predict the quality of a new edit. Quality is predicted using features of the edit itself, the author of the edit, and the article being edited. Through learning to predict quality, we also learn about factors that influence quality. Specifically, we provide analysis of model parameters to determine which features are the most useful for predicting quality.

We advocate using edit quality predictions and information gleaned from model analysis not to place restrictions on editing, but to assist users in improving quality. That is, we aim to maintain a low barrier to participation, as those users not interested in the Wikipedia community can still be valuable contributors (Anthony, Smith, and Williamson

2007). Restrictions might also discourage new users, and drive away users who were drawn to the idea of a openly editable encyclopedia. Consequently, we suggest that the quality models be used to help users focus on predicted quality problems or to encourage participation.

We evaluate the edit quality prediction models and provide analysis using the Spanish Wikipedia. Experiments demonstrate that the models attain better results when given access to content-based features, in addition to features of the contributing user. This suggests that a user-based solution to the Wikipedia quality problem may not be sufficient.

Although we focus on Wikipedia in this paper, we think of this as an instance of a new problem: automatically predicting the quality of contributions in a collaborative environment.

Related Work

Many researchers are skeptical of Wikipedia, as there are reasons to expect it to produce poor quality articles (Denning et al. 2005). Surprisingly, however, a study found that Wikipedia articles are only of slightly lower quality than their counterparts in Britannica, a professionally written encyclopedia (Giles 2005). As a result, Wikipedia has attracted much interest from the research community.

Stvilia et al. (2008) present an overview of the mechanisms used by the Wikipedia community to create and maintain information quality, and describe various categories of quality problems that occur. Roth (2007) analyzes factors that allow Wikipedia to remain viable while other wikis fail. The primary requirements for a viable wiki are quality content and a sufficient mass of users that can maintain it. It is suggested that an overworked user base may result in the abandonment of a wiki. This motivates our work, which aims to provide tools to help users maintain quality.

A controversial aspect of Wikipedia is that any user is allowed to edit any article. However, there is evidence that this openness is beneficial. Riehle (2006) interviews high profile Wikipedia contributors who support the open edit policy, but want more explicit incentive systems and better quality control (again motivating the work in this paper). Anthony et al. (2007) find that both unregistered users with few edits, or "good Samaritans", and registered users with many edits, or "zealots" contribute high-quality edits. Interestingly, as the number of edits contributed increases, quality decreases for unregistered users, whereas it increases for registered users. Although the "good Samaritan" users seemingly make edits without the need for recognition, some registered users are clearly interested in being recognized. Forte and Bruckman (2005) examine the motivation of registered users in Wikipedia and compare their incentives to those in the scientific community. Similarly to researchers, Forte and Bruckman argue that some Wikipedia users want to be recognized by their peers, and gain credibility that will help them to effect change. This suggests that developing better methods for attributing credit to users would benefit Wikipedia.

A reputation system is one way to attribute credit to users. Adler and Alfaro (2007) propose an automatic user reputation system for Wikipedia. In addition to encouraging high-quality contributions, this system can be used to identify potential quality problems by flagging edits made by low reputation users. Similarly to the work in this paper, Adler and Alfaro quantify the quality of a user's edits by observing the community reaction to them in the edit history. When they use their author reputation scores to classify low-quality edits, the resulting precision is fairly low. This is to be expected because users with good intentions but few edits have low reputation. Additionally, a large portion of edits come from unregistered users who have low reputation by default. In the previous paragraph, it was suggested that these users can be beneficial. This motivates a quality prediction model that considers information about the edit itself in addition to information about the user. Although we do not compare directly with the method of Adler and Alfaro, we compare with a quality prediction model that only has access to user features, and find that the addition of edit content features consistently improves performance.

There has also been work that aims to detect quality at the granularity of articles, rather than edits. Wilkinson and Huberman (2007) find that articles with more editors and more edits are of higher quality. Dondio et al. (2006) propose a heuristic method for computing the trustworthiness of Wikipedia articles based on article stability and the collaborative environment that produced the article. Kittur (2007) shows that the number of edits to meta (non-article) pages is increasing, illustrating that more effort is being expended on coordination as Wikipedia grows. Kittur also uses features that quantify conflict to predict whether articles will be tagged *controversial*.

Defining Edit Quality

It has been argued that there is no general definition of information quality, and hence quality must be defined in relation to the community of interest (Stvilia et al. 2008). That is, quality is a social construct. If we are able to observe the operation of the community of interest, however, we can use the actions of the community to quantify quality.

In this paper, we quantify the quality of an edit to an article using implicit feedback from the Wikipedia community. This feedback can be obtained by observing the article edit history, which is openly available. We choose to use implicit feedback, rather than soliciting quality assessments directly, because it allows us to automatically estimate the quality of any contribution. We propose two measures of quality. Both are based on the assumption that edits to an article that are retained in subsequent versions of the article are of high quality, whereas edits that are quickly removed are of low quality. Although this assumption may be locally violated, for example by edits to a current events page, in the aggregate this assumption seems reasonable.

In Wikipedia terminology, a *revert* is an edit that returns the article to a previous version. That is, we say that the jth edit to an article was reverted if the ith version of the article is identical to the kth version of the article, where $i < j < k$. These events often occur when an article is vandalized, or when an edit does not follow the conventions of the article. The first quality measure we propose is simply whether or not an edit was reverted.

A problem with the revert quality judgment is that it only indicates contributions of the lowest quality — those which provide no value (as judged by the community) and are completely erased. We would also like to know about other ranges of the quality spectrum.

Therefore, we define a quality metric which we call *expected longevity*. To do so, we introduce some notation. We denote the ith version of article a as a_i. Each a_i is represented as a sequence of tokens, so that the kth token of the ith version of the article is a_{ik}. Each $a_{ik} = \langle s, i \rangle$, where s is the token text, and i indicates the edit that introduced the token[1]. Let $D(a_i, a_j)$ be the set of tokens that appear in a_j, but not in a_i. Let t_i be the time of the ith edit to article a. We define the expected longevity of edit i, $l(a, i)$, as:

$$l(a,i) = \sum_{j=i+1}^{k} \left(1 - \frac{|D(a_{i-1}, a_i) - D(a_j, a_{j-1})|}{|D(a_{i-1}, a_i)|}\right)\left(t_j - t_i\right),$$

where k is the first edit in which all of the tokens changed or added in edit i have been subsequently changed or deleted. The first parenthesized value in the above equation computes the proportion of the tokens added or changed in edit i that were removed or changed by edit j. Therefore, the expected longevity is the average amount of time before a token added or changed by edit i is subsequently changed or deleted. In some cases, the tokens added or changed by a particular edit are never subsequently changed or deleted. In this case, the unedited tokens are treated as though they were edited by the last edit to the article.

However, there is a potential problem with the above definition. Suppose that the tokens added or changed by edit i are entirely changed or deleted by edit $i + 1$, but edit $i + 1$ is reverted by edit $i + 2$. In this case, we expect $l(a, i)$ to be unfairly small. We handle this problem by ignoring edits that were reverted in the computation of expected longevity.

We note that the expected longevity quality metric (unlike the revert quality metric) is undefined for edits that only delete content. Additionally, we note that including time in expected longevity could be problematic for articles in which edits are very infrequent. We plan to address these issues in future work.

Predicting Edit Quality

We next develop models to predict the quality of new edits. We choose to use machine learning, rather than some handcrafted policy, because it gives statistical guarantees of generalization to unseen data, allows easy updating as new training data becomes available, and may provide more security since the model family, features, and parameters would (ideally) be unknown to attackers.

We choose to learn a single model that can predict the quality of an edit to any Wikipedia article, rather than separate models for each article. This allows the model to learn broad patterns across articles, rather than learn some very specific aspects of a particular article. However, we can still give the model access to article-specific information with this setup.

Predicting whether an edit will be reverted is a binary classification problem. We can model expected longevity directly using regression, but we may not necessarily require precise estimates. In this paper we instead use discrete categories for expected longevity intervals.

Importantly, the models need to be scalable to make learning on Wikipedia-scale data tractable. Therefore, we choose a simple, discriminatively-trained probabilistic log-linear model. Discriminative training aims to maximize the likelihood of the output variables conditioned on the input variables. An advantage of discriminative training is that it allows the inclusion of arbitrary, overlapping features on the input variables without needing to model dependencies between them. We leverage this capability by combining many different types of features of the user, the content of the edit, and the article being edited.

Importantly, we also use aggregate features, which can consider both the edit in question and all other edits in the training data. For example, in addition to a feature for the user of a particular edit, we can include a feature that counts the number of other edits the user contributed in the training data. We provide more discussion of features in the next section.

The probability of a quality label y given a particular edit $\mathbf{x}$ and the set of all edits in the training data $\mathbf{X}$ is

$$p_\lambda(y|\mathbf{x};\mathbf{X}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_i \lambda_i f_i(\mathbf{x}, \mathbf{X}, y)\right),$$

where f_i are feature functions. Although we allow arbitrary features over a single quality label y and all of the input variables in the training data, in this paper we do not model dependencies between the quality labels of different edits. We have some preliminary evidence that accounting for sequential dependencies between the quality predictions of edits to the same article may be beneficial, and plan to pursue this in future work.

We choose parameters $\hat{\lambda}$ for this model that maximize the conditional log likelihood of the training data D, with an additional Gaussian prior on parameters that helps to prevent overfitting. The optimization problem is then

$$\hat{\lambda} = \arg\max_\lambda \sum_{(\mathbf{x},y)\in D} \log p_\lambda(y|\mathbf{x};\mathbf{X}) - \sum_i \frac{\lambda_i^2}{2\sigma}.$$

We choose parameters $\hat{\lambda}$ using numerical optimization.

Features

In this section we describe the feature functions f_i that we use in the quality prediction models. We define content features as the set of all features below except those under the **User** heading.

Change Type features quantify the types of changes the edit makes. We include features for the log of the number of *additions*, *deletions*, and *changes* that the edit makes, as well as the proportions of each of these change types. Additionally, we use features that specify which of change types

[1] We determine correspondences between article versions using a differencing algorithm.

are observed, for example *delete only*, and differences between change type counts, for example the log of *additions - deletions*.

Structural features quantify changes to the structure of the article. Specifically, there are features for *links*, *variables*, *headings*, *images*, and other forms of wiki markup, all concatenated with the change type. For example, one possible feature is *add_link*.

Word features look at the specific words that are added, deleted, or changed by an edit. That is, for a specific word w we have features for w concatenated with a change type, for example *delete_w*. Before adding word features, we strip punctuation and lowercase the text. We also aggregate over the complete history to obtain *low-quality* and *high-quality lexicons*, which are used as features. Finally, we use regular expression features for capitalization patterns, numbers, dates, times, punctuation, and long, repetitive strings.

Article features provide a way to incorporate information about the specific article being edited into the global model. Specifically, we include a feature for each edit that indicates the article to which the edit was made, as well as a feature for the popularity (measured in terms of the log of the number of edits) of the article.

User features describe the author of the edit. We use the *username* of the author (or prefixes of the IP address if the user is unregistered) as a feature, as well as whether or not they are a *registered* user. We identify each registered user as a *bot* or a *human* with a binary feature. We additionally include aggregate features for the log of the number of edits the user has made to any article, the specific article being edited, any meta page, and the discussion page for the specific article being edited. There are also binary features that specify that a user has never edited one of the above types of pages. Finally, there are features that indicate the number of high and low quality edits the user contributed in the training data.

We also include a feature for the log of the epoch time at which the edit was contributed. This feature is helpful because we observe that reverts are becoming more common with time.

An additional set of features we are working to include are based on probabilities of changes under generative models of the articles. Features that quantify the formality or informality of language and features that identify subjectivity and objectivity would also be useful.

Data Processing

We perform experiments using a dump of the Spanish Wikipedia dated 12/02/07. We choose to use the Spanish Wikipedia because the English Wikipedia complete history dump failed for several consecutive months[2], and the authors have some familiarity with Spanish, making it easier to perform error analysis. The Spanish Wikipedia contains over 325,000 articles, and is one of the top 10 largest Wikipedias.

[2]A complete history dump of the English Wikipedia has recently completed successfully. We plan to use it for future experiments.

For the results that follow, we randomly selected a subset of 50,000 articles, each with at least 10 edits. It is common for a single user to make a sequence of edits to an article in a short amount of time. We collapse these sequences (within 1 hour) into a single edit. After collapsing, the total number of edits in this set is 1.6 million. We then tokenize the input so that words and wiki markup fragments are tokens.

To find reverted edits, we look for cases in which article version a_{i-c} is the same as article version a_i, for $2 \leq c \leq C$. This signifies that edits $i-c+1$ through $i-1$ were reverted. This requires $O(Cn)$ comparisons for each article, where n is the number of edits to the article, and C is a constant that specifies the maximum size of the window. In this paper, we use $C = 5$.

A naive implementation of the computation of expected longevity would require $O(n^2)$ runs of a differencing algorithm per article. However, if we maintain a data structure that represents the subsections of the article that were added or changed by each edit, we can do this in linear time. For each article version, we compute its difference from the previous article version using an implementation of longest common subsequence dynamic programming algorithm. We use this information to update the data structure, and the expected longevity of previous edits whose additions and changes were changed or deleted by the most recent edit. We ignore edits that were reverted, so that the expected longevity is a more accurate indicator of how long an edit remains in the article. This requires $O(n)$ runs of the differencing algorithm per article.

The processed data contains millions of features, making learning slow, and causing the model to overfit. As a result, we remove features that occur less than 5 times in the training data.

For these experiments, we use the edits from December of 2006 through May of 2007 for training, and the edits from June 2007 for evaluation. Importantly, the evaluation data is held-out during training, so that we observe the ability of the model to predict the quality of unseen edits, rather than describe the available edits.

Quality Prediction Experiments

In this section, we compare the quality predictions obtained with a model that uses all features, and a baseline model that uses only user features. We compare these models using precision-recall curves for predicting whether an edit is low quality (either reverted or in the lowest longevity category). We present summary statistics of these curves using the maximum F_α measure. The F_α measure is defined as $\alpha pr/(\alpha p + r)$, where p is precision and r is recall. The F_1 measure is then the harmonic mean of precision and recall. The F_2 measure weights recall twice as much as precision, and the $F_{.5}$ measure weights precision twice as much as recall.

We provide results with all three of these statistics because there are arguments for preferring both high recall and high precision systems. A high recall system does not miss many low-quality edits, but may be imprecise in its predictions. A high precision system only flags edits as low-quality

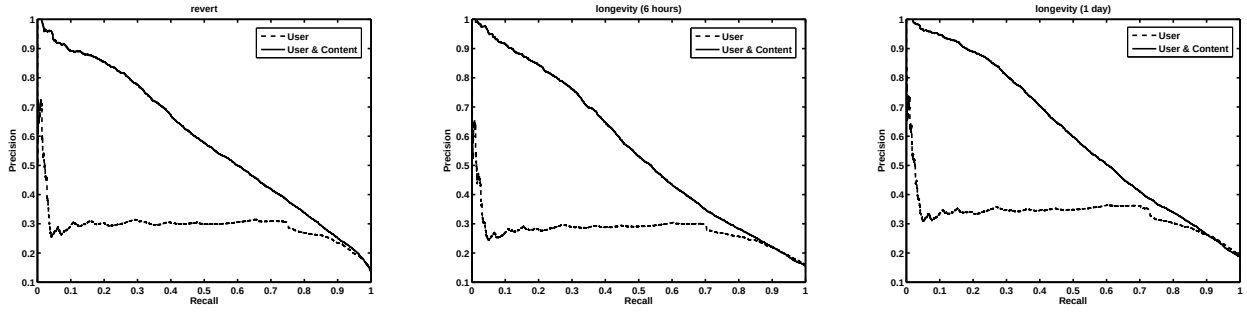

Figure 1: Precision vs. recall curves for models that use only user features and models that additionally use content features to predict the revert, expected longevity (6 hours), and expected longevity (1 day) quality metrics. The model with content features outperforms the user model except in some places at the high recall end of the curve.

features	max F_1	max $F_{.5}$	max F_2
user	0.435	0.382	0.505
+ content	**0.547**	**0.551**	**0.573**

Table 1: Results for predicting reverted edits on the test data.

features	max F_1	max $F_{.5}$	max F_2
user	0.419	0.370	0.483
+ content	**0.518**	**0.538**	**0.535**

Table 2: Results for predicting expected longevity (6 hours).

features	max F_1	max $F_{.5}$	max F_2
user	0.477	0.431	0.535
+ content	**0.550**	**0.567**	**0.569**

Table 3: Results for predicting expected longevity (1 day).

if the model is very confident, but this means that some edits that are actually low quality may be missed.

We first evaluate the quality prediction models on the task of predicting that a new edit will be subsequently reverted. The F_α results are presented in Table 1, while the precision recall curves are presented in Figure 1. We note that the model with access to content features, in addition to user features, performs better in all three F_α comparisons.

We next evaluate quality models on the task of predicting the expected longevity of a new edit. For these experiments we use expected longevity cutoffs such that edits with expected longevity less than 6 hours or 1 day are considered low quality. The results are presented in Tables 2 and 3, and Figure 1. Again, the model with content features performs better under all three F_α measures.

We next aim to understand the poor performance of the user features model. We observe in the user features model precision-recall curves in Figure 1 that there are a very small number of edits for which the precision is high. These edits are contributed by users who often submit low-quality edits. Near recall of 1, the precision dips because edits performed by good users are starting to be classified as low quality. In between, the precision-recall curves are essentially flat, because the users are either unseen during training (33% of the users in June 2007 test data are unobserved in training data), or there is not enough information about them to make a clear decision.

We consider these results promising. The learning task is extremely difficult because the quality metrics are noisy and correctly classifying some edits would require a deep semantic understanding of the article. We also note that there are many other features which would likely improve the results. However, we argue that a model that can, for example, predict whether an edit will be reverted with 80% precision and 30% recall could be useful to Wikipedia users.

Analysis

We now analyze the parameters of the model in order to increase our understanding of the factors that influence edit quality. Some of the most important features for predicting reverts are presented in Table 4 (the important features for predicting longevity are similar). Below, we discuss these and other important features in detail.

- Although unregistered users do contribute 75% of the low-quality edits, they also contribute 20% of all high-quality edits. Therefore, bias against unregistered users results in a system with high recall but low precision.
- Users who previously contributed high or low quality edits tend to continue to submit high and low quality edits, respectively.
- Unregistered users with one or no previously contributed edits often contribute high-quality edits.
- As the number of edits a registered user contributes increases, the quality of their edits increases.
- The percentage of low-quality edits is increasing over time. For example, in October 2005, 8.8% of all edits were reverted, whereas 11.1% of edits in October 2006 were reverted, and 17.8% of all edits in October 2007 were reverted.

↓ NUM USER REVERTED CONTRIBUTIONS
↑ NUM USER HIGH-QUALITY CONTRIBUTIONS
↓ EDIT EPOCH TIME
↓ ADD LOW-QUALITY WORD
↑ REGISTERED USER EDIT COUNT
↓ USER PAGE EDIT COUNT
↑ ADD LINK
↓ ADD ONLY
↓ CHANGE POSITIVE SIZE
↓ CHANGE NEGATIVE SIZE
↑ ADD PUNCTUATION
↓ DELETE LINK
↑ UNREGISTERED USER EDIT COUNT ZERO OR ONE
↑ ADD HIGH-QUALITY WORD
↓ ARTICLE EDIT COUNT

Table 4: The most important features for predicting quality. An ↑ indicates the feature is associated with high quality, whereas a ↓ indicates low quality.

- Articles that are more popular, where popularity is measured in terms of the number of edits, receive a higher percentage of low-quality edits.
- The adding of a link, heading, or other structural element tends to indicate high quality, whereas changing or deleting a structural element indicates low quality.
- Adding punctuation indicates high quality.
- There exist lists of words that tend to indicate high and low-quality edits.
- Large changes in the size of the article, whether a result of additions or deletions, indicate low quality. This suggests that the Wikipedia users who maintain articles are reluctant to allow major changes.
- Surprisingly, edits contributed by users who have edited the article in question many times are often low-quality. This is likely a result of edit wars.

Example Application: Improved Watch List

Wikipedia users can be notified of changes to articles by joining the article's *Watch List*. We suggest an improved Watch List that prioritizes edits according to the confidence of quality predictions. In addition to notifying users on the list, we can also seek out other qualified users to address quality problems. To ensure that the user is knowledgeable about subject of the article, we can leverage work on the reviewer assignment problem. We can determine reputable authors by using a simple reputation system based on the quality metrics. An advantage that the quality prediction models afford is that we can avoid the so-called "ramp-up" problem with author reputation. Typically, a reputation system cannot assign a meaningful reputation score to a new or unregistered user, or incorporate recent contributions, because time is needed for the community to assess them. With the aid of a quality prediction model, we can use estimated quality values for new edits, allowing us to have a meaningful reputation estimate immediately.

Conclusion

We have used the implicit judgments of the Wikipedia community to quantify the quality of contributions. Using relatively simple features, we learned probabilistic models to predict quality. Interestingly, a model that has access to features of the edit itself consistently outperforms a model that only considers features of the contributing user. Through analysis of the parameters of these models, we gained insight into the factors that influence quality. Although we have focused on Wikipedia, we think of this as an instance of a new problem: automatically predicting the quality of contributions in a collaborative environment

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010, and AFRL #FA8750-07-D-0185. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

Adler, B. T., and de Alfaro, L. 2007. A content-driven reputation system for the wikipedia. In *WWW*, 261–270.

Anthony, D.; Smith, S. W.; and Williamson, T. 2007. The quality of open source production: Zealots and good samaritans in the case of wikipedia. Technical Report TR2007-606, Dartmouth College, Computer Science.

Denning, P.; Horning, J.; Parnas, D.; and Weinstein, L. 2005. Wikipedia risks. *Commun. ACM* 48(12):152–152.

Dondio, P.; Barrett, S.; Weber, S.; and Seigneur, J. 2006. Extracting trust from domain analysis: A case study on the wikipedia project. *Autonomic and Trusted Computing* 362–373.

Forte, A., and Bruckman, A. 2005. Why do people write for wikipedia? incentives to contribute to open-content publishing. In *GROUP 05 Workshop on Sustaining Community: The Role and Design of Incentive Mechanisms in Online Systems*.

Giles, J. 2005. Internet encyclopaedias go head to head. *Nature* 438:900–901.

Kittur, A.; Suh, B.; Pendleton, B. A.; and Chi, E. H. 2007. He says, she says: conflict and coordination in wikipedia. In *CHI*, 453–462.

Riehle, D. 2006. How and why wikipedia works: an interview with angela beesley, elisabeth bauer, and kizu naoko. In *Proceedings of the International Symposium on Wikis*, 3–8.

Roth, C. 2007. Viable wikis: struggle for life in the wikisphere. In *Proceedings of the International Symposium on Wikis*, 119–124. ACM.

Stvilia, B.; Twidale, M. B.; Smith, L. C.; and Gasser, L. 2008. Information quality work organization in wikipedia. *JASIST* 59(6):983–1001.

Wilkinson, D. M., and Huberman, B. A. 2007. Cooperation and quality in wikipedia. In *Proceedings of the International Symposium on Wikis*, 157–164.