On Measuring the Quality of Wikipedia Articles

Gabriel De La Calzada and Alex Dekhtyar

ABSTRACT

This paper discusses an approach to modeling and measuring information quality of Wikipedia articles. The approach is based on the idea that the quality of Wikipedia articles with distinctly different profiles needs to be measured using different information quality models. We report on our initial study, which involved two categories of Wikipedia articles:"stabilized" (those, whose content has not undergone major changes for a significant period of time) and"controversial" (the articles, which have undergone vandalism, revert wars, or whose content is subject to internal discussions between Wikipedia editors). We present simple information quality models and compare their performance on a subset of Wikipedia articles with the information quality evaluations provided by human users. Our experiment shows, that using special-purpose models for information quality captures user sentiment about Wikipedia articles better than using a single model for both categories of articles.

INTRODUCTION

In the past few years, due to significant expansion of its content, and due to high ranking its articles receive from web search engines, Wikipedia [29] has become the go-to location for a wide range of information for millions of Internet users. Running on an open source MediaWiki [21] platform, Wikipedia has been the trailblazer for the open content collaborative model of information collection and presentation. While one of the goals Wikipedia pursues, outlined in the words of its founder Jimmy Wales as "...a world in which every single person on the planet is given free access to the sum of all human knowledge" [30, 22]: collecting and providing access to a large body of information, parallels the goals of other encyclopedia, Wikipedia takes a distinctly different approach to achieving this task.

Where traditional encyclopedia rely on domain experts to produce content, and use formal and rigorous editorial and peer review process to validate it before allowing public access to the information, Wikipedia allows anyone to edit existing entries and create new ones. It relies on the collective wisdom of many readers-cum-editors to prevent, and, if necessary, fix, erroneous, false, poorly presented or simply inappropriate content. This approach to content creation in Wikipedia has been subject to numerous arguments between its critics [27, 20] and its defenders [10, 28].

Research on the quality and reliability of Wikipedia has largely concentrated on comparative analysis of Wikipedia articles and articles from traditional encyclopedia. The results [31] have been quite varied. Studies found both high-quality, full and well-written content, as well as incomplete and, at times, poorly written articles [28, 1].

This leads to the following observation. Due to its high visibility on the web, Wikipedia plays an important role in the information collection and dissemination, and overall, enjoys the reputation of an easy-to-access, easy-to-understand, and reasonably reliable information source [6].

At the same time, the content within Wikipedia itself is quite diverse in its quality and reliability. Individual Wikipedia articles range from well-thought-out and thoroughly edited 100+ Kb essays (like, for example some of the Wikipedia's featured articles, i.e., articles that appear on its front page), to simplistic 2-3 sentence article stubs. Users accessing Wiki-pedia content encounter different articles and thus, are exposed to information of varying quality.

In addition to the comparative quality and reliability assessment studies mentioned above [6, 28, 1], a new research direction, concentrating on direct assessment and/or estimation of information quality of Wikipedia articles has emerged recently [33, 25, 13, 26]. These approaches analyze the text of the individual articles, as well as the rich meta-data that Wikipedia makes available about the articles, such as the edit history, internal discussions, and the actual change history, to determine how "good", informative and/or reliable a specific article is.

The research described in this paper continues this avenue of investigation and introduces two important aspects into it. First, our work concentrates on comparing the performance of our information quality models to the opinions of Wikipedia users. While comparative Wikipedia studies [6, 28, 1]

usually rely on teams of experts and peer review process to assess and compare Wikipedia articles (thus applying the traditional encyclopedia validation methodology to the study of Wikipedia), we observe, that the open nature of Wikipedia makes opinions of individual users reading its articles for information just as important. Therefore, discovering how Wikipedia users determine for themselves the quality and reliability of individual Wikipedia articles is, in our view, an important question.

To advance our discovery process, we observe, that by their nature, and, often, by their history, Wikipedia articles (and their topics) can fit a number of different profiles (which, for simplicity, we refer to as categories for the rest of the paper). The content of some articles on Wikipedia has reached its saturation point some time ago (e.g., the article on Benjamin Franklin), while the nature of some other articles (e.g., an article on Barack Obama) dictates frequent significant changes/updates to account for the new developments. Yet other articles (e.g., the article on religion), due to their topics become subjects to controversy, vandalism, revert wars and heated debates among the editors.

Others works [33, 25, 13, 26] have concentrated on using a single method to assess the quality of any Wikipedia article. In this paper we hypothesize that (a) users of Wikipedia use different criteria for assessing quality of articles from different categories and (b) different information quality models, when used in concert, better predict the opinion of Wikipedia users about the quality of the information they read.

This paper describes our initial study. For this study we chose to look at two categories of Wikipedia articles, which we termed stabilized (see Section 3.1) and controversial (see Section 3.2). For each category of articles, we developed a specific information quality assessment method. To determine how well our methods, in concert, or by themselves, estimate the quality of articles we conducted a controlled study in which a number of human participants was shown a variety of Wikipedia articles and asked to evaluate the their information quality. Our results showed that when the quality of stabilized and controversial articles was estimated using their respective evaluation methods, the overall prediction error (as compared to the mean user opinion) was lower, than when only one method (either for controversial article quality estimation, or for stabilized article estimation) was used for all studied articles.

The rest of this paper is organized as follows. Section 2 outlines the background information necessary for our study, and discusses some related work on information quality of on-line resources. In Section 3 we describe the properties of two categories of Wikipedia articles: stabilized and controversial, and present the information quality models we developed for each of the two categories. In 4 we describe the experimental study we conducted to determine human opinion about information quality of some Wikipedia articles. We describe the experimental setup and data collection procedures and report on the results of comparing human opinion to the information quality predictions delivered by our models.

2. BACKGROUND AND RELATED WORK

2.1 Wikipedia

Wikipedia is an example of a globally accessible online encyclopedia, where anyone can participate in the preservation of knowledge. This approach lies in stark contrast to traditional sources of information such as encyclopedias. Wikipedia is an open content project, meaning anyone with an Internet connection can modify or create an article. This openness even allows anonymous, non-registered users to make significant contributions to existing articles. Wikipedia's philosophy is that as the community works together on content, the content becomes more reliable over time. Consequently, articles found on Wikipedia are never "finished" as modifications are continuously made. In addition, openness is traded for the lack of formal peer review [12]. Although Wikipedia has come a long way, there is no formal mechanism for a peer review by subject matter authorities. It is also known that many articles do not cite their primary sources [12].

The open content ideal behind Wikipedia makes vandalism and misinformation possible, and self interested parties have taken advantage of this in the past. Wikipedia has temporarily banned access to Wikipedia from government domains in response to a rising trend of defacement of political candidates

[19]. Political operatives have been reported to modify Wikipedia entries to make a certain candidates appear strong or weak.

Our study has used MediaWiki API (http://www.mediawiki.org/wiki/API) to retrieve a variety of meta-data about Wikipedia articles. MediaWiki [21] is the open source wiki software platform used by Wikipedia. The MediaWiki API for Wikipedia is publicly available (http://en.wikipedia.org/w/api.php) and is accessible through PHP via specially crafted URIs. The parameter list of such a URI determines the specifics of the query. With the MediaWiki API, it is possible to query information from articles, login into the MediaWiki application, post changes to articles and to obtain meta-data (such as the revision history) for Wikipedia articles.

2.2 Comparative Reliability Studies

Ever since Wikipedia's introduction, numerous studies comparing Wikipedia to traditional sources of knowledge have been conducted, as documented in [31]. The majority of these studies compare Wikipedia to an authority such as traditional peer-reviewed sources or a team of experts.

The results from a number of studies suggest that Wikipedia suffers from major errors of omission. The study conducted by [5] analyzed Wikipedia articles for seven top Western philosophers. These articles where then compared to a consensus list of themes acquired from various works in philosophy. From this comparison, it was found that the Wikipedia articles on average covered only 52% of the list of themes. However, no errors were found in the content of these articles.

Similarly, in [11] a research team analyzed 80 Wikipedia articles on drugs. They found that the articles often missed important information and a small number of factual errors.

On the other hand, a number of studies suggest that Wikipedia is no worse or if not, better than existing peer reviewed sources of information. In [7], 50 Wikipedia articles were compared to their counterparts in a German encyclopedia "Brockhaus Enzyklop¨adie" [4]. Results showed that on average Wikipedia articles were more accurate, complete and up to date, while the Brockhaus articles were judged to be more clearly written. A number of other studies [6, 2, 3, 8] compared the content of selected Wikipedia articles to other encyclopedia, including Encarta and Encyclopedia Britanica. These comparisons did judge Wikipedia to be less reliable than the traditional encyclopedia.

Comparative studies help "calibrate" the public perception of the quality and reliability of Wikipedia in general. However, these studies involve tiny (and not always representative) portions of the Wikipedia. Additionally, while Wikipedia itself relies on achieving quality through article evolution, comparative studies mimic the validation procedures used by conventional encyclopedia. Our work described in the paper uses human assessment of article quality, but relies on peer assessment rather than expert reviews.

2.3 Work On Information Quality

This section overviews another approach to Wikipedia article quality assessment: direct estimation.

The authors in [9] measure quality of individual article contributions as the percentage of a contributor's text in the current version of the article. The authors found that dedicated registered users that make many contributions, and anonymous low contribution users generate the highest quality contributions. Similarly, in [18] the authors discovered correlations between a Wikipedia article's quality and the categories of its authors.

In [14], the authors proposed and evaluated four different quality models: Naive, Basic, PeerReview, and ProbReview. In the Naive model, the quality of an article is directly proportional to the number of words contained in that article. The Basic model co-opts the HITS framework [16], which determines the hub and authority scores of web pages, to the problem of estimation of the quality of Wikipedia articles. The higher the authority of the authors of an article, the higher is the quality of that article. Authority of a user is based on the quality of the articles that user has authored. Both article quality (Qi)and user authority (Aj), enforce each other as shown below:

$$Q_i = \sum_j c_{ij} A_j \qquad A_j = \sum_i c_{ij} Q_i \qquad (1)$$

The third model, PeerReview, identifies a separate quality of each word in an article. Quality of a word is based on the authority of the user who authored the word, and the authority of any user who reviewed the word. This approach, thus, rewards words that survived multiple review cycles. The authority of a user is based on the quality of the words the user has authored or reviewed. The sum of word qualities belonging to a single article is interpreted as the overall quality of the article. The PeerReview model is summarized in the equation below:

$$q_{ik} = \sum_{w_{ik} \xleftarrow{A} u_j \cup w_{ik} \xleftarrow{R} u_j} A_j \qquad (2)$$

Here, q_{ik} is the quality of the k^{th} word wik in article ai, wik + uj the set of words authored by user uj, and wik + uj the set set words reviewed by user uj .

The ProbReview model assumes that a user who submits a revision to an article does not necessarily review every word in that submission. For example, a user skimming through an article might notice that certain statistics are missing from the article and submits a revision which contains the original article content in addition to the new statistics. In this case, the new statistics were authored, however the remaining content wasn't reviewed. The ProbReview model is a modification of the PeerReview model. It takes into account the probability that a user submitting a revision has reviewed a word in a document. Equations 3, 4, and ?? model quality in the ProbReview model. Function Prob determines the probability that user uj reviewed the word wik. The intuition behind this function is that when a user authors content of an article, that user is more likely to review content located closer to the newly authored content.

$$q_{ik} = \sum_j f(w_{ik}, u_j) A_j \qquad (3)$$

$$A_j = \sum_{i,k} f(w_{ik}, u_j) q_{ik} \qquad (4)$$

$$f(w_{ik}, u_j) = \begin{cases} 1 & \text{if } w_{ik} \xleftarrow{A} u_j \\ Prob(w_{ik} \xleftarrow{R} u_j) & \text{otherwise} \end{cases} \qquad (5)$$

A study by Lih[17] focuses on the "reputation" of an article. Lih's model assumes that the more reputable an article, the higher its quality. Reputation in this context is the amount of collaborative work that went into the authoring of an article. Instead of focusing on the actual content of an article for quality assessment, Lih's methodology focuses only on article's metadata. Specifically, the model relies on information found directly in an article's revision history. In this model, rigor is defined as the total number of revisions to a particular article. The assumption is that the more revisions an of an article, the deeper the treatment of the subject and higher scrutiny on the content. Diversity is defined as the total number of unique users contributing to an article. The assumption is that more unique contributors means more voices and different points of view on the subject of a given article. Articles whose rigor and diversity are both above the media are considered to be of high quality.

Zeng et al. [33] propose a quality model which focuses on the trustworthiness of an article. This model recognizes that articles evolve over time, and thus their trustworthiness evolves over time. An article that was trustworthy a month ago might not be trustworthy today. The trust of an article is based on the trust of the previous version of the article, the trust of the current author, and any insertions or deletions. Trust is a continuous number ranging from [0, 1], where a trust of 0 is most untrustworthy while a trust of 1 is most trustworthy. This model uses a dynamic bayesian network to model trust.

In [25], the authors used machine learning to construct an automated quality assessment system. The authors identified six quality classes of articles from worst to best: stub, B-article, good article, A-article, and featured article. The quality class of an article was predicted using a classifier based on the maximum entropy model. The classifier made use of over 50 features which fell into one of the following four categories: length measures, part-of-speech usage, web-specific features, and readability metrics.

In [13], Dalip utilized the same machine learning approach as [25] to assess Wikipedia article quality.

However, Dalip treats the problem of automatic quality assessment of Wikipedia articles as a regression analysis problem and uses a support vector regression classifier to solve it [26]. The classifier uses the quality classes from [25]. Thus, an article predicted as ”stub” is assigned stub quality while an article predicated as ”Featured-Article” is assigned featured article quality.

Our approach to evaluating the information quality of a Wikipedia article is similar to the approaches described in this section. We use a variety of information about an article to develop models for predicting its quality. However, whereas all work described above uses one quality assessment/prediction model for all Wikipedia pages, we investigate a two-tier approach in which we first determine a broad category of a given article, and then use category-specific quality prediction model to compute the information quality estimate. Additionally, we validate our models and our approach empirically, by investigating, how well they predict the quality assessments made by casual Wikipedia visitors.

3. QUALITY MODELS

We propose a two-step approach to evaluating and/or prediction the information quality of Wikipedia articles. First, we separate Wikipedia articles into a number of *categories*, based on their history and the nature of their topics. Unlike [13] and [25], which split Wikipedia articles *horizontally* by the perceived quality, we split the articles *vertically*: articles belong to the same category if they exhibit similar properties, not if they are of similar quality. On the second step, we develop a quality prediction model for articles within each category and apply it to estimating the information quality of the articles.

Overall, we have established six categories of Wikipedia articles: (1) *stabilized articles*, (2) *controversial articles*, (3) *evolving articles*, (4) *list*, (5) *stub* and (6) *disambiguation page*. This list is not exhaustive: other categories can be defined in a manner described below. For our initial study presented here, we elected to concentrate on two categories of articles: stabilized and controversial. In Sections 3.1 and 3.2 we define these categories and construct article quality models for them.

3.1 Stabilized Article

Informally, a *stabilized Wikipedia article* is one that has more or less ”caught up” with the total knowledge of the topic and is considered to be complete content-wise. Stabilized article topics, typically, refer to events, people, notions, etc., that no longer change over time. The changes to these types of articles are mostly either ”maintenance” revisions, such as those made by automated bots to update the categories of articles, or the reverts of a random vandalism attack. Since a stabilized article is supposed to be complete content-wise, we expect in general to find significant accuracy of the content relative to the total topic knowledge.

To model the quality of stabilized articles Wikipedia’s ”featured articles” can serve as quality benchmarks. Wikipedia features some of the better-written complete articles on its front page on a rotating basis. Wikipedia’s policy mandates that featured articles must be stable. Their content may not be subject to an ongoing edit war and ”... does not change significantly from day to day, except in response to the featured article process” [32]. As such, featured articles are essentially what other stabilized articles “aspire” to be.

Our proposed quality model uses a collection of article features listed in table 1. It is based on the intuition that all of these article features except for length, can be considered as necessary building blocks for an article. For instance, images, references, citations, paragraphs, and links are all hallmarks of a quality article. However, too much or too little of these building blocks can cause an article to be over-or under- developed.

There was no rigorous effort to determine the best features for stabilized articles. The stabilized model is intended to be a simple model as part of a more complicated article classification scheme (see section 3.3). Thus, we choose features which appeared reasonable for a stabilized article and which were simple to extract.

Featured articles serve as benchmarks of quality. The model postulates that when a stabilized Wikipedia article has the exact same proportion of characteristics to the ”typical” featured article then the

effect of article length its quality at its strongest. However, as article's characteristics deviate from those of a"typical" featured article, then the influence of article length diminishes.

This model requires a sample of featured articles from Wiki-pedia. The sample is interpreted as a collection of mixture components of a mixture model. Within this mixture model are six mixture components derived from the sample set of featured articles. These mixture components are the Gaussian probability density functions for logarithm of length, citation density, internal link density, external link density, image count density, and section count density. A single mixture component *i* is computed as follows:

$$C_i(article) = \frac{1}{\sqrt{2\pi\sigma^2}}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Here, where μ is the mean value for the component and σ is the standard deviation. For example, the "length in bytes" component represents a Gaussian probability density function for the length in bytes of a sample of featured articles. Within each mixture component, the standard deviation σ is multiplied by a "forgiveness factor". This forgiveness factor controls how strict or lenient the component is. A default factor of 2 is used in this model.

The quality of an article is represented as the normalized sum of mixture components:

$$q(article) = \frac{\sum C_i(article)}{\sum Cmax_i}$$

3.2 Controversial Article

Controversial articles are articles whose topic or content are subject to a range of opinions. Wikipedia editorial policy requires neutral point of view narratives, but Wikipedia editors are human, and, on occasion, their biases make it into the text of the articles they edit, intentionally or unintentionally. When other editors detect such biases and disagree with them, the article may become a subject to controversy. Some articles are inherently controversial due to the nature of their topic and content (for example the article on Religion) Other articles 'may be going through a controversial phase due to certain attention-grabbing current event or other circumstances. Controversial articles are often the target of vandalism and act as a battleground for revert wars. Historically, a controversial article can be characterized by large number of reverts due to vandalism and revert wars and a high number of anonymous contributions.

We model the quality of controversial articles by taking into account their revision history. Our quality model is similar to the mixture model used for stabilized articles, however it uses different article features, shown in Table 2, to represent controversial articles. Each component is a Gaussian probability distribution and the final quality score takes into account all mixture components as shown in the formulas below:

$$C_i(article) = \frac{1}{\sqrt{2\pi\sigma^2}}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$q(article) = \frac{\sum C_i(article)}{\sum Cmax_i}$$

3.3 Categorizing the Articles

Before the quality model for either stabilized or controversial articles can be applied to a given Wikipedia article, we must first determine if an article is stabilized or controversial (or if it belongs to a different category). We achieve this using supervised learning (classification) techniques. In particular, for each article category, we develop and train a classifier. Given a Wikipedia article, finding its quality is a two-step process. First, the article's features are extracted, and are run against a battery of classifiers (only two for the experiments described in this paper). When a classifier in the series positively classifies the target article, a quality model corresponding to the classifier type is applied to the article. For the case where a target article is classified as positive by multiple classifiers in the series, *the average of outputs for each applied quality model* is used as the final score of the target article. Finally, for the case where

the target article is not positively classified by any classifier in the series, the stabilized model of article quality is utilized as the final score. (In the experiments described in this paper, we only consider articles that were positively classified by at least one of our classifiers.)

Each classifier was trained from a dataset of 96 Wikipedia articles. This dataset was manually chosen to include a mix of each article type described in the previous sections. Class labels for this dataset were manually assigned. A number of supervised learning algorithms provided by WEKA [23] were then utilized to build classifiers for this dataset. Among these algorithms, the one which provided the best results was chosen as the algorithm for the final classifier. In this case, the sequential minimal optimization (SMO) [24, 15] learning algorithm for training a support vector machine classifier was chosen. Using leave-one-out cross validation, the precision (percentage of correct predictions) and recall (coverage percentage) for the SMO classifiers for stabilized and controversial articles are shown in Tables 3 and 4.

4. EVALUATION

Prior research on information quality in Wikipedia, described briefly in Section 2.3 ([14, 17, 33, 25, 13]) approaches computing quality of an article in a uniform manner: for each proposed method the quality of any and all articles is estimated in exactly the same way. In contrast, our approach is to recognize that there may be inherent differences in how the quality of different Wikipedia articles should be estimated. We use different techniques and/or information for articles which belong to different "categories."

Our pilot study was designed to test the hypothesis, that using separate models to compute quality estimates for articles of different types leads to higher accuracy. We selected two categories of articles described in Section 3, stabilized and controversial articles. As the means of validation, we elected to compare the predictions of our models to the opinions of casual Wikipedia users. As such, the study described below pursued two main questions: (1) do information quality models for stabilized and controversial articles adequately predict human opinion of the stabilized and controversial articles respectively? and (2) does using two models to predict information quality lead to more accurate predictions, then using a single information quality model for all articles?

4.1 Quality

In most prior work, the "golden standard" for information quality is the evaluations of experts [31, 7, 5, 11]. This standard has a clear advantage: it is as objective, as it gets. It also has a clear disadvantage: the majority of Wikipedia articles is observed by casual readers in search of information, and their perception of quality is different than that of the experts. In our study we choose a quality evaluation approach that parallels Wikipedia's content creation approach. Just as an individual user may provide incorrect information, the quality assessment of an individual reader may be skewed. However, a combined quality assessment obtained from multiple casual readers will provide a clear idea of what a reader should expect from an article.

4.2 Study

To test our information quality models we conducted an experimental study in which participants read a variety of Wikipedia pages and ranked their information quality. The study involved 247 Cal Poly students who were enrolled during the Fall 2009 quarter in an array of courses (both major courses and service courses) offered by the Computer Science department.

To conduct the study, we have created a dataset consisting of 100 Wikipedia articles and used the versions of those articles offered to the readers on October 20, 2009. We used the "frozen" version of each article instead of the current version to ensure that all subjects who observed/read a specific Wikipedia article in our study accessed exactly the same content. Among the 100 articles, 51 were selected by us while the remainder of the articles were chosen randomly, using Wikipedia's "return a random article" feature. We chose to select a subset of articles directly to ensure that articles of each

category we were interested in were present in the dataset. We also chose some articles to ensure the presence in the dataset of articles about topics that are both well-known to study participants (e.g., "Cal Poly") as well as rather unknown (e.g., "Choi Jai-Soo"). We applied stabilized and controversial articles classifiers obtained from using WEKA's [23] SMO algorithm [24, 15] implementation. Table 7 shows that of our 100 articles, 50 were classified as stabilized, 29 as controversial, 10 as both, and 31 as neither.

Each study participant, via a specially designed on-line software tool (see Figure 1) received access to eight pages from our sample'. The survey software maintained information on the number of times each article has been assigned to study participants. When a new user accessed the software, a list of eight different articles was randomly drawn from our dataset, with the probability distribution which granted article(s) with the fewest number of assignments the highest chance of being selected. Use of this procedure lead to each article being shown to roughly the same number of participants. In our study, each article was viewed and assessed by 18—20 participants.

For each page, we asked the participant to (a) read it, (b) evaluate its information quality and (c) specify the level of familiarity with the topic of the article. Participants could evaluate the information quality on a scale from 1 to 5. The familiarity was evaluated on a scale from 1 to 3. The full scales are shown in Tables 5 and 6.

4.3 Measures

At the conclusion of the survey, we had accumulated a number of information quality and user confidence ratings for each article A provided by individual participants. We used the average user rating $\bar{q}_u(A)$ to represent user opinion about each article. Of the 100 articles in our dataset, we considered only the 69 with were classified as stabilized or controversial in the analysis described below. For each article A from this list, we computed two scores q_s (A) and q_c (A) using the stabilized and controversial information quality models described in Section 3 respectively, and the score qmix(A), which was computed as follows:

$$q_{mix}(A) = \begin{cases} q_s(A) & \text{A is stabilized, not controversial;} \\ q_c(A) & \text{A is controversial, not stabilized;} \\ \frac{q_s(A)+q_c(A)}{2} & \text{A is stabilized and controversial.} \end{cases}$$

Further, we computed the errors of prediction $\delta_s(A) = q_s(A) - \overline{q_u}(A)$ and $\delta_c(A) = q_c(A) - \overline{q_u}(A)$ and $\delta_{mix}(A) = q_{mix}(A) - \overline{q_u}(A)$ for the stabilized, controversial models and mixed models respectively. To test our hypotheses, we compared the average overall prediction errors for each model: $\overline{\delta_s} = \frac{1}{|S|}\sum_{A \in S} \delta_{mix}(A)$ and $\overline{\delta_{mix}} = \frac{1}{|S|}\sum_{A \in S} \delta_{mix}(A)$ (here, S is a set of articles over which the prediction error is computed).

4.4 Results

Figure 2 depicts the results of the stabilized method prediction, i.e., q_s, (Section 3.1) on articles classified as stabilized plotted vs. the average reader opinion $\bar{q}_u$. Figure 3 shows q_s, the controversial scores, plotted vs. the average reader opinion $\bar{q}_u$ for articles classified as controversial. Table 8.(a) shows the $\bar{\delta}_s$, δ_c and δmix for the set of stabilized articles, the set of controversial articles and the set consisting of both stabilized and controversial articles.

4.5 Analysis

Our first question was whether the two models we selected in this paper to measure the information quality of stabilized and controversial articles respectively were sufficiently accurate to make their further study meaningful. As seen from Figures 2 and 3, when applied to stabilized articles only, the stabilized model showed clear positive correlation with the average opinion. The correlation between the controversial model predictions and the reader opinion for controversial articles appears to be somewhat less pronounced (as seen on Figure 3), however excluding a few outliers, there is still a distinct positive

correlation. In fact, Table 8.(a) shows that the average error for controversial articles scored by the controversial model is 0.103 (with standard deviation of 0.0989): lower than 0.127, the average error for stabilized articles scored by the stabilized model (with standard deviation of 0.0781). Both methods achieve an error of 10–12%, which, given the simplicity of the model suggests to us, that the methods are reasonably accurate.

The second question we considered, and the main question of our study, is whether the two-step approach to article quality prediction is justified. First and foremost, as seen from Table 8 the stabilized model performs better than the controversial model on stabilized articles (mean error of 0.127 vs. mean error of 0.201), while the controversial model outperforms the stabilized model (mean error of 0.103 vs. mean error of 0.124) on controversial articles. The Student T-test (Table 8.(b)) shows that the first of these differences is statistically significant at the 5% significance level, while the second is not.

More importantly, the mixed scoring model outperforms the other two models (mean error of 0.116 vs. mean errors of 0.127 and 0.175). Here, Student T-Test shows that the difference between mean errors for the mixed and controversial models is statistically significant at the 5% significance level, and the difference for the mixed and stabilized models is not. The T-test statistics are shown in Table 8.(b).

5. CONCLUSIONS AND FUTURE WORK

We presented the first in a series of results comparing a variety of quality models for Wikipedia articles with the opinions of casual Wikipedia readers. In this paper, we were able to confirm the key assumption behind our approach to measuring information quality: that quality of articles of different "type" should be computed using different means. We also introduced a new approach to validating models: validating quality estimates against the combined opinion of multiple casual Wikipedia readers, rather than against opinions of individual experts.

We plan to explore this topic further, and study the following questions. First, we are interested in validating other article quality models [17, 33, 25, 13] versus the casual reader opinion. Second, we plan to expand our study to include other categories of Wikipedia articles, such as evolving articles mentioned in Section 3. Third, we want to conduct a comparative study of a variety of quality models for each category. Last, but not least, we will investigate what affects the quality scores assigned to articles by casual readers. Our experimental study produced a variety of data (some of which had to be left out of this paper for space considerations) which can shed more light on how non-expert readers evaluate quality of information on-line.

6. ACKNOWLEDGMENTS

We would like to thank Franz Kurfess, Clark Turner, Kevin O'Gorman, Mark Hutchenreuther, and Kurt Voelker: Cal Poly professors who kindly allowed us to administer the experimental study in their classes.

7. REFERENCES

[1] Can you trust wikipedia? The Guardian, October 2005.
[2] Internet encyclopedias go head to head. Nature, December 2005.
[3] Survey of wikipedia accuracy and completeness. http //bpastudio.csudh.edu/fac/lpress/wikieval, 2006.
[4] Brockhaus Enzyklop¨adie. Brockhaus, Mannheim, 21 edition, 2007.
[5] Wiki-philosophizing in a marketplace of ideas Evaluating wikipeida's entries on seven great minds. Available at SSRN http //ssrn.com/abstract=978177, April 2007.
[6] Wikipedia uncovered. PC Pro Magazine, August 2007.
[7] Wikipedia: Wissen f¨ar alle. Stern, December 2007.
[8] Wikipedia vs. encyclopedia A question of trust? are online resources reliable or should we stick to traditional encyclopedias? Techradar.com, April 2008.

[9] D. Anthony, S. Smith, and T. Williamson. Explaining quality in internet collective goods: Zealots and good samaritans in the case of wikipedia. Electronically.
[10] T. B. Barry X. Miller, Karl Helicher. I want my wikipedia! Library Journal, April 2006.
[11] K. Clauson, H. Polen, M. K. Boulos, and J. Dzenowagis. Scope, completeness, and accuracy of drug information in wikipedia. Ann Pharmacother, 2008.
[12] P. Denning, J. J. Horning, D. Parnas, and L. Weinstein. Wikipedia risks. Communications of the ACM, 48(12), 2005.
[13] D. Hasan Dalip, M. Andr´e Gon¸calves, M. Cristo, and P. Calado. Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In JCDL '09: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, pages 295–304, New York, NY, USA, 2009. ACM.
[14] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. Measuring article quality in wikipedia: models and evaluation. In CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pages 243–252, New York, NY, USA, 2007. ACM.
[15] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, 1999.
[16] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604—632, September 1999.
[17] A. Lih. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In 5th International Symposium on Online Journalism, 2004.
[18] J. Liu and S. Ram. Who does what: Collaboration patterns in the wikipedia and their impact on data quality. In 19th Worksho on Information Technologies and Systems, 2009.
[19] S. McCaffrey. Political dirty-tricksters are using wikipedia. The Mercury News, April 2006.
[20] R. McHenry. The faith-based encyclopedia. Tech Central Station, Nocember 2004.
[21] MediaWiki. http //mediawiki.org.
[22] R. Miller. Wikipedia founder jimmy wales responds. Slashdot, 2004.
[23] U. of Waikato. Weka 3: Data mining software in java. http://www.cs.waikato.ac.nz/ml/weka.
[24] J. Platt. Machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning. MIT Press, 1998.
[25] L. Rassbach, T. Pincock, and B. Mingus. Exploring the feasibility of automatically rating online article quality.
[26] V. N. Vapnik. The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA, 1995.
[27] S. Waldman. Who knows? The Guardian, October 2004.
[28] D. Wiegand. Entdeckungsreise. digitale enzyklop¨adien erkl¨aren die welt. c't, March 2007.
[29] Wikipedia. English Edition, http //en.wikipedia.org.
[30] Wikipedia. http //en.wikipedia.org/wiki/Jimmy Wales.
[31] Wikipedia. http //en.wikipedia.org/Wikipedia reliability.
[32] Wikipedia. Wikipedia: Featured article criteria. http://en.wikipedia.org/wiki/Wikipedia Featured article criteria, November 2009.
[33] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness. Computing trust from revision history. In PST, page 8, 2006.

8. Figures and Images

Pos.	Feature Name	Description
1	Log Length	Base 10 logarithm of article length in bytes
2	Citation Density	Citations per article length
3	Internal Link Density	Internal links per article length
4	External Link Density	External links per article length
5	Image Count Density	Images per article length
6	Section Count Density	Sections per article length

Table 1: Article features used in modeling quality of stabilized articles.

Feature Name	Description
Avg. Number of Reverts	Average number of reverts in the article's revision history
Revisions Per Registered User	Average revisions per registered authors
Revisions Per Anonymous User	Average revisions per anonymous authors
Percentage of Anonymous Users	Percentage of anonymous authors

Table 2: Article representation in the controversial model

Class	Precision	recall
false (not stabilized)	0.800	0.784
true (stabilized)	0.761	0.778
Weighted Avg.	0.782	0.781

Table 3: Stabilized Classifier Evaluation

Class	Precision	Recall
false (not controversial)	0.972	0.920
true (controversial)	0.760	0.905
Weighted Avg.	0.925	0.917

Table 4: Controversial Classifier Evaluation

Value	Description
1.0	Fail. Not ready for public consumption
1.5	-
2.0	Poor. Requires more revisions
2.5	-
3.0	Average. Serves its purpose
3.5	-
4.0	Good. Nearly complete
4.5	-
5.0	Great. Requires minor/no further revisions

Table 5: Survey Article Quality Scale

Value	Description
1.0	The subject is new to me
1.5	-
2.0	I have decent understanding
2.5	-
3.0	I'm an expert at the subject

Table 6: Survey Confidence Scale

Article Type	Total
All	100
Stabilized	50
Controversial	29
Both	10
Neither	31

Table 7: Survey Article Classification

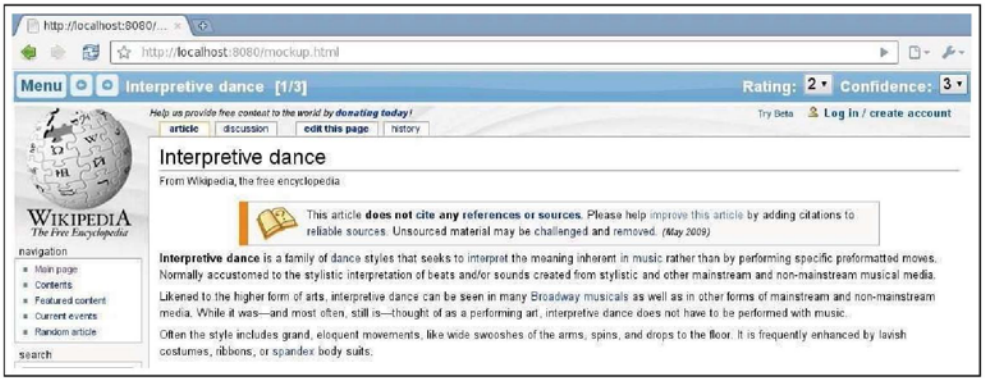

Figure 1: Main Survey Interface

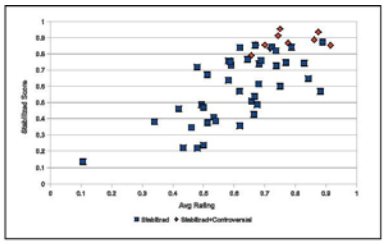

Figure 2: Avg Rating vs Stabilized Score

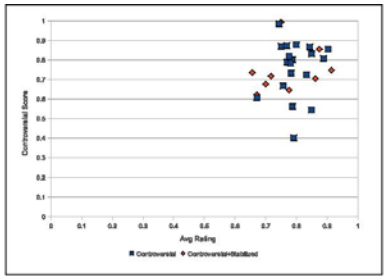

Figure 3: Avg Rating vs Controversial Score

	N	Stabilized Model	Controversial Model	Mixed Model
Stabilized Articles	50	Mean: 0.127 StDev: 0.0781	Mean: 0.201 StDev: 0.145	Mean: 0.119 StDev: 0.0802
Controversial Articles	29	Mean: 0.124 StDev: 0.0814	Mean: 0.103 StDev: 0.0989	Mean: 0.0979 StDev: 0.0934
Both Categories	69	Mean: 0.127 StDev: 0.811	Mean: 0.175 StDev: 0.141	Mean: 0.116 StDev: 0.0880

(a)

Error: Mix vs. Stabilized (all articles)
DF: 135
T-Value: -0.7639
Mix < Stabilized: P-Value = < 0.2231
Error: Mix vs. Controv. (all articles)
DF: 113
T-Value: -2.95
Mix < Controv.: P-Value = < 0.0019
Error: Stabilized vs Controv. (Stabilized Articles)
DF: 75
T-Value: -3.1897
Stabilized < Controv.: P-Value = < 0.0010
Error: Controv. vs Stabilized (Controv. Articles)
DF: 52
T-Value: -0.8678
Controv. < Stabilized: P-Value = < 0.1947

(b)

Table 8: Mean absolute prediction errors (left) and T-test results for mixed model vs. stabilized and controversial models (right).