Statistical Measure of Quality in Wikipedia

Sara Javanmardi
University of California Irvine
sjavanma@ics.uci.edu

Cristina Lopes
University of California Irvine
lopes@ics.uci.edu

ABSTRACT

Wikipedia is commonly viewed as the main online encyclopedia. Its content quality, however, has often been questioned due to the open nature of its editing model. A high–quality contribution by an expert may be followed by a low–quality contribution made by an amateur or a vandal; therefore the quality of each article may fluctuate over time as it goes through iterations of edits by different users. With the increasing use of Wikipedia, the need for a reliable assessment of the quality of the content is also rising. In this study, we model the evolution of content quality in Wikipedia articles in order to estimate the fraction of time during which articles retain high–quality status. To evaluate the model, we assess the quality of Wikipedia's featured and non–featured articles. We show how the model reproduces consistent results with what is expected. As a case study, we use the model in a CalSWIM mashup the content of which is taken from both highly reliable sources and Wikipedia, which may be less so. Integrating CalSWIM with a trust management system enables it to use not only recency but also quality as its criteria, and thus filter out vandalized or poor–quality content.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Computer–supported cooperative work, web-based interaction

General Terms

Collaborative Authoring, Groupware

Keywords

Wikipedia, Wiki, Crowdsourcing, Web 2.0

1. INTRODUCTION

Web 2.0 is the second generation of the web that emphasizes crowdsourcing, the process of outsourcing a task to a large group of people, in the form of an open call [1]. Using wiki technology, Wikipedia has become the largest crowdsourcing project and the main online encyclopedia [2]. It has been suggested that wiki technology can harness the Internet for science; "Wikinomics" is a recent term that denotes the art and science of peer production when masses of people collaborate to create innovative knowledge resources [3]. Because of its open editing model –allowing anyone to enter and edit content– Wikipidia's overall quality has often been under question. While it is difficult to measure Wikipedia's overall quality in a definitive way, two studies have tried to assess it manually by comparison of Wikipedia articles to their parallel articles in other reputable sources [4, 5]. *Nature* magazine's comparative analysis of forty–two science articles in both Wikipedia and the Encyclopedia Britannica showed a surprisingly small difference; Britannica disputed this finding, saying that the errors in Wikipedia were more serious than the Britannica errors and that the source documents for the study included the junior versions of the encyclopedia as well as the Britannica year books[1].

Since Wikipedia is a highly dynamic system, the articles are changing very frequently. Therefore, the quality an article is a time–dependent function and a single article may contain high– and low–quality content in different spans of its lifetime. In this paper we develop an automated measure to estimate the quality of article revisions throughout the entire English Wikipedia. Using this statistical model, we follow the evolution of content quality and show that the fraction of time that articles are in a high–quality state has an increasing trend over time. We show that non–featured articles tend to have high–quality content 74% of their lifetime and this is 86% for featured articles. Furthermore, we show that the average article quality increases as it go through various edits.

To address the problem of content quality in a real–world application, we have developed a scientific mashup called CalSWIM [6]. CalSWIM is an information and management tool designed both as a public forum for exploring watersheds and as a web location for professionals to acquire data. Leveraging the power of "crowdsourcing", CalSWIM provides a specialized view of Wikipedia's articles related to Water Resources. To smooth out the quality challenges for the content fetched from Wikipedia, we are integrating the mashup with a trust management system that can automatically assign reputation to the contributors of the wiki articles and estimate the quality of their content. This feature helps CalSWIM users interested in Wikipedia articles

1st Workshop on Social Media Analytics (SOMA '10), July 25, 2010, Washington, DC, USA.
Copyright 2010 ACM 978-1-4503-0217-3 ...$10.00.

[1] http://bit.ly/cLDpXO

have access to the most recent, high–quality revision of the article (as opposed to Wikipedia's normal practice of showing merely the most recent revision).

The remainder of this paper is organized as follows: Section 2 describes related work. In Section 3, we explain how the data was collected. Section 5 shows how content quality is modeled. Section 5 provides a brief overview of CalSWIM and describes the reputation management system for Wikipedia. Finally, Section 6 draws some conclusions and provides some direction for future investigation.

2. BACKGROUND & RELATED WORK

In the open editing model of Wikipedia, users can contribute anonymously or with untested credentials. As a consequence, the quality of Wikipedia articles has been a subject of widespread debate. For example, in late 2005, American journalist John Seigenthaler publicly criticized Wikipedia because of a collection of inaccuracies in his biography page, including an assertion that he was involved with the assassination of former U.S. President John F. Kennedy[2]. Apparently the inaccuracies remained in Wikipedia for 132 days. Because there is no single entity taking responsibility for the accuracy of Wikipedia content, and because users have no other way of differentiating accurate content from inaccurate content, it is commonly thought that Wikipedia content cannot be relied upon, even if inaccuracies are rare [7].

To overcome this weakness, Wikipedia has developed several user–driven approaches for evaluating the quality of its articles. For example, some articles are marked as "featured articles". Featured articles are considered to be the best articles in Wikipedia, as determined by Wikipedia's editors. Before being listed here, articles are reviewed as "featured article candidates", according to a special criteria that takes into account: accuracy, neutrality, completeness and style[3]. In addition, Wikipedia users keep track of articles that have undergone repeated vandalism in order to eliminate it and report it [4]. However, these user–driven approaches cannot be scaled and only a small number of Wikipedia articles are evaluated in this way. For example, as of March 2010, only 2,825 articles (less than 0.1%) in English Wikipedia are marked as featured. Another difficulty of the user–driven evaluations is that Wikipedia content is, by its nature, highly dynamic and the evaluations often become obsolete rather quickly.

As a result, recent research work involves the automatic quality analysis of Wikipedia [8, 9, 10, 7, 11, 12, 13, 14, 15, 16]. Cross [7] proposes a system of text coloring according to the age of the assertions in a particular article; this enables Wikipedia users to see what assertions in an article have survived after several edits of the article and what assertions are relatively recent and thus, perhaps, less reliable. Adler *et al.* [17] quantify the reputation of users according to the survival of their edit actions; then they specify ownerships of different parts of the text. Finally, based on the reputation of the user, they estimate the trustworthiness of each word. Javanmardi *et al.* in [8] present a robust reputation model for wiki users and show that it is not only simpler but also more precise compared to the previous work.

Other research methods try to assess the quality of a Wikipedia article in its entirety. Lih [12] shows that there is a positive correlation between the quality of an article and the number of its editors as well as the number of its revisions. Liu et. al. [13] present three models for ranking Wikipedia articles according to their level of accuracy. The models are based on the length of the article, the total number of revisions and the reputation of the authors, who are further evaluated by their total number of previous edits. Zeng *et al.* [15] compute the quality of a particular article revision with a Bayesian network from the reputation of its author, the number of words the author has changed and the quality score of the previous version. They categorize users into several groups and assign a static reputation value to each group, ignoring individual user behavior.

Stvilia et. al. [14] have constructed seven complex metrics using a combination of them for quality measurement. Dondio *et al.* [11] have derived ten metrics from research related to collaboration in order to predict quality. Blumenstock [10] investigates over 100 partial simple metrics, for example the number of words, characters, sentences, internal and external links, etc. He evaluates the metrics by using them for classifications between featured and non–featured articles. Zeng *et al.*, Stvilia *et al.* and Dondio *et al.* used a similar method which enables the evaluation results to be compared. Blumenstock demonstrates, with an accuracy of classification of 97%, that the number of words is the best current metric for distinguishing between featured and non–featured articles. These works assume that featured articles are of much higher quality than non–featured articles, and recast the problem as a classification issue. Wohner and Peters [16] suggest that, with improved evaluation methods, these metrics–based studies enable us to determine the accuracy of various submissions. Studying the German Wikipedia, they believe that a significant number of non–featured articles are also highly accurate and reliable. However, this category includes a large number of short articles. Their study of German Wikipedia from January 2008 shows that about 50% of the articles contain less than 500 characters, and thereby they assume that some short non–featured articles are of high quality, since their subject matter can be briefly but precisely explained.

In addition, we and others [18, 16] assume that when an article is marked as featured and is displayed on its respective wiki pages, it attracts many more web users for contributions and demands more administrative maintenance. Wohners and Peters' investigation on German Wikipedia[16] reveals this assumption to be true. For example, over 95% of all articles are edited with greater intensity, once they are marked as featured. Wilkinson and Huberman [18], in a similar study on English Wikipedia, show that featured articles gain an increase in the number of edits and editors after being marked as featured. According to these observations, the accuracy of the classification in the related work ([15, 14, 11]) will be valid only if featured articles are considered before they are marked as featured.

3. METHOD AND DATASET

Most of the content analysis research on the evolution of articles (like those enumerated in Section 2 and our own work) require the full text of all revisions of articles. We have monitored the publicly available English Wikipedia dumps[5]

[2] http://bit.ly/4Bmrhz

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

[4] http://bit.ly/dy3t1Y

[5] http://download.wikimedia.org/enwiki/

since early 2006 with the last successful dump released in October 2007. Because of the exponential growth of Wikipedia, all of the history dumps have failed since then. Since the last dump data set is quite out–dated, we created a more recent data set which is now publicly available[6]. We used the Wikipedia API[7] to get the full text of all the submitted revisions in the history of Wikipedia. The API has a limit of 50 revisions per request and, since these types of requests are not frequent, the chance of having a cached version is slim which makes the process of fetching data expensive. On average, it takes more than one second for the server to send back the result for each request. In addition, we needed to compare the text of subsequent revisions in order to extract the edits made in a revision. This process is also computationally expensive. In order to maintain a reasonable processing speed and still remain polite to Wikipedia servers, we used a cluster of ten nodes which downloaded and processed the whole history of English Wikipedia from July through August 2009. A master node assigned articles to client nodes and waited for them to download and process the article history and send back the extracted statistics.

As of May 2010, English Wikipedia contains about 3.3M articles [8]. However, some portion of these articles are isolated stubs that are not referenced by any other article. In our analysis, we used Crawler4j[9] to crawl the entire English Wikipedia and extract a list of articles accessible through links on the English Wikipedia home page[10]. We also ignored articles that were redirected to other articles. We ended up with a set of 2.2M articles. Then we downloaded the revisions of these articles through Wikipedia API which resulted in 130M revisions[11].

4. MEASURING ARTICLE QUALITY

A concept closely related to information trust is information quality. Kelton *et al.*[19] describe trust as playing a key role as a mediating variable between information quality and information usage. Hence, trust can be seen as an assessment of information quality upon which the decision to use the information is based [20]. In this work, we measure an article's quality as an indicator of its trustworthiness.

Since Wikipedia is a dynamic system, the articles can change very frequently. Therefore, the quality of articles is a time–dependent function and a single content may contain high– and low–quality content in different periods of its lifetime. The goal of our study is to analyze the evolution of content in articles over time and estimate the fraction of time that articles are in high–quality state.

In our analysis of the evolution of the content quality in Wikipedia articles, we divide revisions to low– and high–quality revisions. Based on this assumption, an article can be in *low quality* ($q = 0$) or *high quality* ($q = 1$) states. In order to assess the quality q of a revision, we take into account two factors: the reputation of the author and whether this revision has been reverted in one of the subsequent revisions or not.

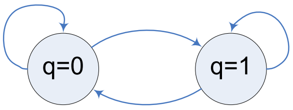

Figure 1: Transitions between high quality and low quality states

The reputation of a user can be viewed as the probability of him producing a high–quality contribution. This probability is computed by methods developed in [8]. The heuristic behind this reputation assessment is that high–quality contributions tend to survive longer in wiki articles. This heuristic is also supported by other work [21, 9]. Assume that user i has inserted $N_i(t)$ tokens in the system before time t and $n_i(t)$ of these tokens are not deleted yet. At the time t, he inserts $c_i(t)$ new tokens where $g_i(t)$ of them remain in the wiki article, and the rest are deleted by other users. Reputation of user i is updated based on the following formula:

$$R_i^+(t) = \max\left(0, \frac{n_i(t) + g_i(t) - \sum_{d=1}^{p_i(t)} R_{j(t_d)} e^{-\alpha(\triangle r)}}{N_i(t) + c_i(t)}\right) \tag{1}$$

where $R_{j(t_d)}$ is the reputation of the deleter at the time of deletion, $p_i(t)$ is the number of deleted tokens, and $\triangle r$ is the number of revisions submitted between insertion and deletion of the tokens.

When used as a classifier, the model produces an area under the ROC curve of 0.98. Furthermore, we assess the reputation predictions generated by the models on other users, and show that the models can be used efficiently for predicting user behavior in Wikipedia. The effectiveness and efficiency of the model and its comparison with related work is discussed in [8].

As Figure 1 suggests, submission of a new revision can keep the state of the article or move it to the other state. If the revision is reverted later in the article history, we consider the new state of the article to be $q = 0$. Otherwise, if the reputation of the author of that revision is r, then with probability of r the new revision will be $q = 1$ and with probability of $1 - r$ the new revision will be $q = 0$.

With all these elements in place, we define $Q(T)$ as the ratio of high quality revisions submitted for the article up to time T:

$$Q(T) = \sum_{i=1}^{n} q(t_i)/n \tag{2}$$

where $q(t_i)$ is the quality of the revision submitted at time t_i and n is the total number of revisions up to time T. Figure 2 shows the distribution of $Q(T)$ for both all featured articles and a non–featured articles. While the average of $Q(T)$ is relatively high for both featured and non–featured articles, it is higher for featured articles –74% vs. 65%.

To estimate the proportion of time during which an article is in a high–quality state, we also define the duration $QD(T)$ by:

$$QD(T) = \frac{\sum_{i=1}^{n}(t_{i+1} - t_i)q(t_i)}{T - t_1} \tag{3}$$

[6] http://nile.ics.uci.edu/events-dataset-api/
[7] http://en.wikipedia.org/w/api.php
[8] http://en.wikipedia.org/wiki/Special:Statistics
[9] http://code.google.com/p/crawler4j/
[10] http://en.wikipedia.org/wiki/Main_Page
[11] The dataset is publicly available at http://nile.ics.uci.edu/events-dataset-api/

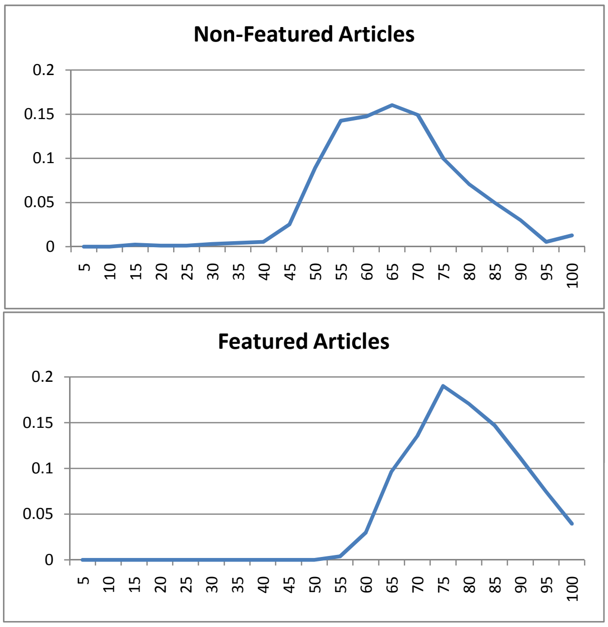

Figure 2: Distribution of Q(T) for featured and non–featured articles

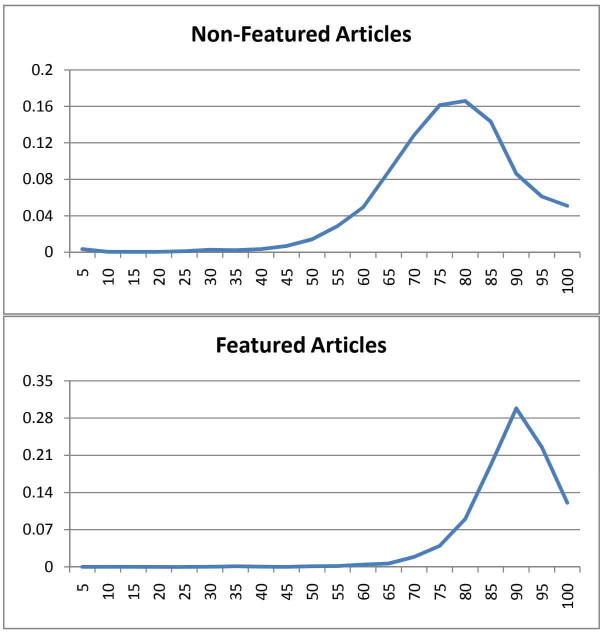

Figure 3: Distribution of QD(T) for featured and non–featured articles

Number of Articles	$20,824$
Number of Registered Users	$101,465$
Number of Anonymous Users	$302,324$
Number of Revisions by Registered Users	$1,236,642$
Number of Revisions by Anonymous Users	$581,804$
Average Reputation of Registered Users	0.6967
Average Reputation of Anonymous Users	0.4202

Table 1: Properties of the Dataset

The distribution of $QD(T)$ for both featured and non–feature articles are shown in Figure 3. Figure 4 also shows the average and standard deviation of $Q(T)$ and $QD(T)$ for both featured and non–featured articles. Featured articles on average contain high–quality content 86% of the time. Interestingly, this value increases to 99% if we only consider the last 50 revisions of the articles. The same statistics for non–featured articles show that they have high–quality content 74% of the time. The difference between the averages of $Q(T)$ and $QD(T)$ suggests that typically low–quality content has short life span. This result is consistent with other studies reporting the rapid elimination of vandalism in Wikipedia [22, 23, 24]. For example, [24] reported that about one third to one half of the systematically inserted fictitious claims in Wikipedia are corrected within 48 hours.

5. CASE STUDY: CALSWIM

The concept of scientific mashups is gaining popularity as the sheer amount of scientific content is scattered over different sources, such as databases or public websites. A variety of mashup development frameworks exist, but none fully address the needs of the scientific community. One limitation of scientific mashups is the issue of trust and attribute; especially when the content comes from collaborative information repositories where the quality of such content is unknown. We have developed a scientific mashup called CalSWIM [6], where the quality of the content fetched from Web 2.0 can be assessed. CalSWIM is an information and management tool designed both as a public forum for exploring watersheds and as a web location for professionals to acquire data. Leveraging the power of "crowdsourcing", CalSWIM provides a specialized view of Wikipedia's articles related to Water Resources. To resolve some of the quality challenges for the articles fetched from Wikipedia, we have integrated the mashup with a trust management system that can automatically assign reputation to the contributors of the wiki articles and estimate quality of the content as explained in Section . This feature helps CalSWIM users interested in Wikipedia articles have access to the most recent, high–quality revision of the article (as opposed to Wikipedia's normal practice of showing merely the most recent revision).

Having evaluated user reputations, we can then rank the recent revisions of an article according to the trustworthiness of their contributors. Then, it is possible to suggest the latest reliable revision of an article to the user. To evaluate the effectiveness of this idea, we calculated the reputation of users contributing to the water–related Wikipedia articles extracted in CalSWIM. Table 1 shows the properties of the dataset.

Our study of the entire English Wikipedia in September 2009 shows that the average reputation for good users (i.e. users who contribute high–quality content) is 82% while, for vandals it is 22% [8]. If we use the same settings here and assume that users who average more than 82% contribute high–quality content and users who average less than 22% contribute low quality content, we can estimate the percentage of high quality content in the most recent revisions of articles. Table 2 summarizes our results. When considering only the last revisions of articles ($n = 1$), about 73% of them are of high–quality, and 1% are of low–quality. When considering the last five revisions of articles ($n = 5$), we found

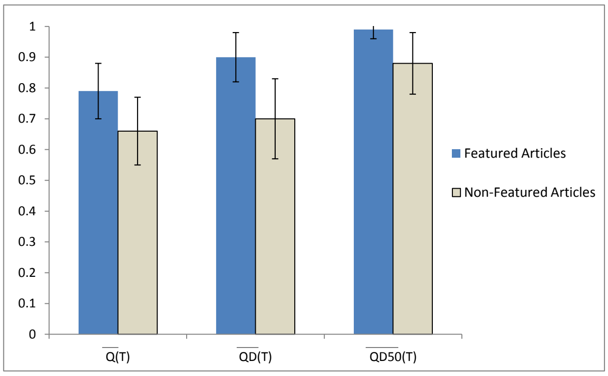

Figure 4: Average article quality for featured articles and non–featured articles. Quality is assessed by the average and the standard deviation of Q, QD, and $QD50$ for featured and non–featured articles. For each article, Q is the ratio of high–quality revisions. QD is the amount of time that an article spends in its high–quality state computed over its entire lifetime. $QD50$ is the value of QD when only considering the last 50 revisions of the article.

	$n = 1$	$n = 2$	$n = 3$	$n = 5$
Rep > 0.82	72.61%	86.28%	92.48%	96.93%
Rep < 0.22	1.0%	0.12%	0.05%	0.01%

Table 2: Percentage of articles with high reputation and low reputation users in their last n revisions. When $n > 1$, results are based on the maximum of the reputation of users contributing the last n revisions.

that for almost 97% of the articles, at least one revision had been submitted by a high reputation user. Therefore, it is more beneficial to show this revision to users rather than merely the most recent one. Figure 5 shows the full distribution of reputation of the users contributing to the last five revisions of the articles.

6. DISCUSSION AND FUTURE WORK

Wikipedia is a highly dynamic environment and the quality of its articles can change over time as they go through iterations of edits by different users. In this work, we modeled the quality of an article's revision using mainly the reputation of its editor and showed that non–featured articles tend to have high–quality content 74% of the time, while featured articles average 86%. Furthermore, we showed that the average article quality increases as it go through edits while its standard deviation decreases.

Our initial study on the application of the trust model in our CalSWIM mashup shows that the measuring quality in addition to mere recency is much more beneficial to the user when he is fetching content from Wikipedia. However, it is important to note that assessing the quality of content based on a contributor's reputation does have its limitations:

- Data sparsity: for a considerable number of users in Wikipedia, we do not have enough information for an accurate assessment of reputation. The model that we employed to evaluate a user's reputation is based on his edits and how others reacted to them. Therefore, in cases in which a user is new to the system, we do not have a stable reputation estimate for him.
- Anonymity: a significant number of users contribute to Wikipedia articles anonymously and they are only identified by their IP addresses. However, there is a loose correspondence between the IP addresses and the real–world users.
- Expertise: the quality of a user's contribution depends on his expertise on that particular topic. Having only one reputation value may not be a perfect indicator of the quality of his contributions on different topics. In the case of CalSWIM we tried to alleviate this problem by estimating the reputation of users based soley on their contributions to water–related articles.

In addition to the above limitations, there is no guarantee that users will not change their behavior in the future. Thus, a user who has contributed high quality content in the past, might contribute low quality content in the future. In addition, when a new user comes to the article and contributes high quality content, the system sacrifices freshness for trustworthiness, only because it does not have an accurate estimate of the user's reputation. This problem becomes worse for articles that are updated less frequently. In the case of our CalSWIM mashup, some articles get updated very infrequently. The average timespan between submission of the last two revisions of articles is 29 days. However, our study on Wikipedia featured articles shows that the update rate for an article increases significantly as it gains more visibility [25]. According to this observation, our conjecture is that mashups like CalSWIM can help these articles gain more visibility and thereby enjoy more frequent updates.

To overcome the limitations caused by inaccurate user reputation, in future work we aim at processing the changes done in newly submitted revisions of an article to see if it is vandalistic or not. Inspired by [26], we will categorize Wikipedia vandalism types and build a statistical language models, constructing distributions of words from the revision history of Wikipedia articles. As vandalism often involves the use of unexpected words to draw attention, the fitness (or lack thereof) of a new edit, when compared with language models built from previous revisions, may well indicate that an edit is the product of vandalism. One of the main advantages of this technique is that it is extendable, even to other Web 2.0 domains such as blogs.

7. ACKNOWLEDGMENTS

This work has been partially supported by NSF grant OCI-074806.

8. REFERENCES

[1] M. Waldrop, "Editorial. big data: Wikiomics," *Nature News*, no. 455, pp. 22–25, 2008.

[2] J. Zittrain, *The Future of the Internet–And How to Stop It.* Yale University Press, 2008.

[3] D. Tapscott and A. Williams, *Wikinomics: How Mass Collaboration Changes Everything.* Penguin Group, 2006, pp. 70–77.

[4] J. Giles, "Internet encyclopaedias go head to head," *Nature*, pp. 438:900–901, December 2005.

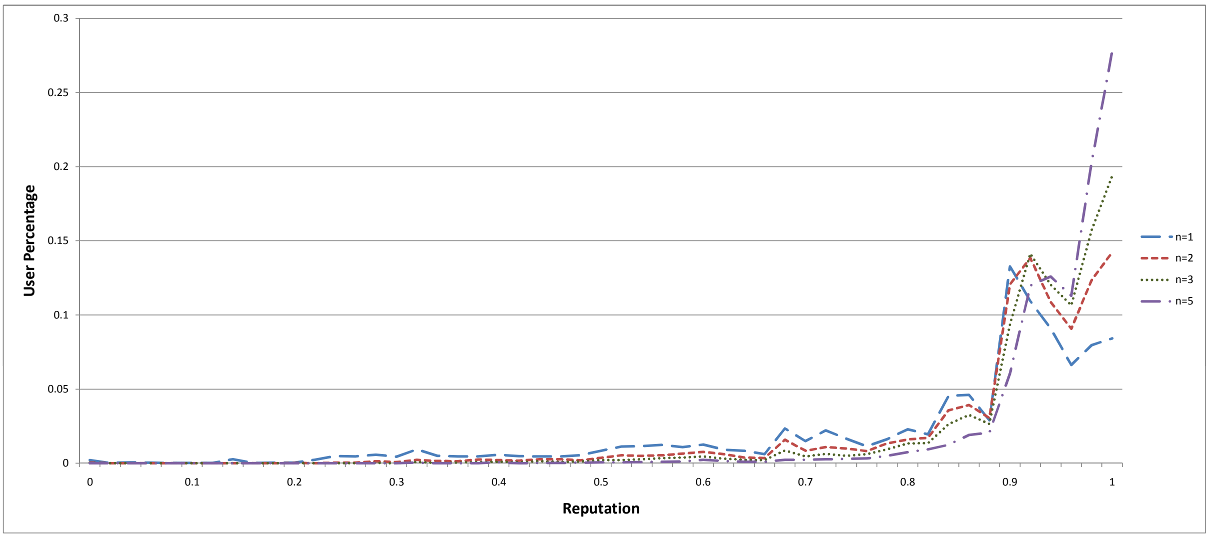

Figure 5: Distribution of user reputation for the last n revision of water–related articles in Wikipedia

[5] T. Chesney, "An empirical examination of wikipedia's credibility," *FirstMonday*, vol. 11, no. 11, 2006.

[6] Calswim mashup. [Online]. Available: http://nile.ics.uci.edu/calswim/

[7] T. Cross, "Puppy smoothies: Improving the reliability of open, collaborative wikis," *First Monday*, vol. 11, no. 9, September 2006.

[8] S. Javanmardi, C.Lopes, and P.Baldi, "Modeling user reputation in wikipedia," *Journal of Statistical Analysis and Data Mining (accepted)*, vol. 3, no. 2, pp. 126–139, March 2010.

[9] B. T. Adler and L. de Alfaro, "A content-driven reputation system for the wikipedia," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM, 2007, pp. 261–270.

[10] J. E. Blumenstock, "Size matters: word count as a measure of quality on wikipedia," in *WWW '08: Proceedings of the 17th international conference on World Wide Web*. ACM, April 2008, pp. 1095–1096.

[11] P. Dondio and S. Barrett, "Computational trust in web content quality: A comparative evalutation on the wikipedia project," *Informatica: An International Journal of Computing and Informatics*, vol. 31, no. 2, pp. 151–160, 2007.

[12] A. Lih, "Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource," in *Proceedings of the 5th International Symposium on Online Journalism*, April 2004.

[13] H. Liu, E. Lim, H. Lauw, M. Le, A. Sun, J. Srivastava, and Y. A. Kim, "Predicting trusts among users of online communities: an epinions case study," in *EC'08: Proceedings of the 9th ACM conference on Electronic commerce*. ACM, 2008, pp. 310–319.

[14] L. S. B. Stvilia, M.B. Twidale and L. Gasser, "Assessing information quality of a community-based encyclopedia," in *In Proceedings of the International Conference on Information Quality*, November 2005, pp. 442–454.

[15] H. Zeng, M. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness, "Computing trust from revision history," in *Proceedings of the 2006 International Conference on Privacy, Security and Trust*, October 2006.

[16] T. WŽhner and R. Peters, "Assessing the quality of wikipedia articles with lifecycle," in *WikiSym '09 Proceedings of the 2009 International Symposium on Wikis*. ACM, October 2009.

[17] B. Adler, L. d. A. K. Chatterjee, M. Faella, I. Pye, and V. Raman, "Assigning trust to wikipedia content," in *WikiSym '08: Proceedings of the 2008 international symposium on Wikis*, May 2008.

[18] B. A. H. D. Wilkinson D, "Assessing the value of cooperation in wikipedia," *First Monday*, vol. 12, no. 4, 2007.

[19] K. Kelton, K. R. Fleischmann, and W. A. Wallace, "Trust in digital information," *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, no. 3, pp. 363–374, 2008.

[20] T. Lucassen and J. M. Schraagen, "Trust in wikipedia: how users trust information from an unknown source," in *Proceedings of the 4th workshop on Information credibility*, April 2010, pp. 19–26.

[21] M. Hu, E. Lim, A. Sun, H. Lauw, and B. 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*. ACM, 2007, pp. 243–252.

[22] F. Viégas, M. Wattenberg, and K. Dave, "Studying cooperation and conflict between authors with history flow visualizations," in *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004, pp. 575–582.

[23] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi, "He says, she says: conflict and coordination in wikipedia," in *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2007, pp. 453–462.

[24] P. D. Magnus, "Early response to false claims in

wikipedia," *First Monday*, vol. 13, no. 9, September 2008.

[25] S. J. Y. Ganjisaffar, , C. Lopes, and P. Baldi, "Statistical measure of the effectiveness of the open editing model of wikipedia," in *CWSM Data Challenge*, May 2010.

[26] R. Lopes and L. Carriço, "Using language models to detect wikipedia vandalism," in *WSDM*, 2010.