

Modeling User Reputation in Wikis

Sara Javanmardi, Cristina Lopes and Pierre Baldi

Bren School of Information and Computer Sciences,

University of California, Irvine, USA

{sjavanma, lopes, pfbaldi}@ics.uci.edu

Abstract

Collaborative systems available on the Web allow millions of users to share information through a growing collection of tools and platforms such as wikis, blogs, and shared forums. By their very nature, these systems contain resources and information with different quality levels. The open nature of these systems, however, makes it difficult for users to determine the quality of the available information and the reputation of its providers. Here, we first parse and mine the entire English Wikipedia history pages in order to extract detailed user edit patterns and statistics. We then use these patterns and statistics to derive three computational models of a user's reputation. Finally, we validate these models using ground-truth Wikipedia data associated with vandals and administrators. When used as a classifier, the best 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 all three models can be used efficiently for predicting user behavior in Wikipedia.

Keywords: Wiki, Reputation, Reliability, Wiki Mining, Wikipedia, Web 2.0

1 Introduction

The last few years have seen a substantial growth in user-generated Web content. Simple editing interfaces encourage users to create and maintain repositories of shared content. Online information repositories such as wikis, forums and blogs have increased the participation of the general public in the production of web content through the notion of social software [1, 2, 3].

Online information repositories, especially in the form of wikis, are widely used on the Web. Wikis are originally designed to hide the association between a wiki page and the authors who have produced it [4]. The main advantages of this feature are: (a) it eliminates the social biases associated with group deliberation, thus contributing to the diversity of opinions and to the collective intelligence of the group, and (b) it directs authors towards group goals, rather than individual benefits [5]. One of the key characteristics of wiki software is its very low-cost collective content creation, requiring only a regular web browser and a simple markup language. This feature makes wiki software a popular choice for content creation projects where minimizing overhead is of high priority; especially in creating new or editing already existing content. For this reason, these platforms can be used in web-based collaborative content management systems for scientific purposes such as team-based research, and e-learning [6, 7, 8].

The most well-known example of a public collaborative information repository is Wikipedia, which has a traffic rank of six worldwide¹. Usually people trust user-generated content in Wikipedia for learning purposes or decision making without validating its information [9]. For these aims, the highly desirable properties of wikis or other similar social software technologies —openness, ease-of-use, and

¹According to traffic report by Alexa.com in December 2009.

decentralization— can also have disruptive consequences on society. Open wikis can easily be associated with poor-quality information, and often fall prey to malicious or misleading content editing [10].

Online communities use trust/reputation management components to facilitate cooperative user behavior [11]. In general, trust management systems seek two main goals: (a) to assist users in rating products or other users for better decision making, and (b) to provide an incentive for better user behavior resulting in improved future performance [12, 13]. In the context of wikis, reputation management systems are suggested as a social rewarding technique that motivates users to participate actively in sharing knowledge [14]. In addition, these systems can assist administrators for automatic detection of high/low reputation users to promote/demote the access rights.

Reputation can be defined as the opinion (more technically, a social evaluation) of the public toward a person, a group of people, or an organization [15]. *Trust* is one user’s belief in another user’s capabilities, honesty and reliability based on his own direct experiences. In online communities, there are two notions of trust: individual-to-individual trust and individual-to-technology trust [16]. eBay and online banking are examples of these two categories, respectively. In Wikis, we have a combination of these trust/reputation relationships; individuals need to have trust in content that is collaboratively created by other individuals. Authors also need to have trust in other authors collaborating with them to create/edit content. For example, one of the obstacles experts who collaborate with Wikipedia face is the lack of guarantee that an inexpert/vandal user will not tamper with their contributed content [17]. Therefore, the trustworthiness of content is tightly linked with the reputation of the author.

In this work, we focus on estimating the reputation of wiki users based on their edit patterns. We parse and mine entire English Wikipedia history pages in order to extract detailed user edit patterns and statistics. We use these patterns and statistics to derive three computational models of users’ reputation. The main contribution of this work is to accurately infer reputation of users through these mathematical models that are simpler than models previously proposed. With these models, reputation criteria are based on the users’ actions in the system, rather than the explicit, or subjective judgment used in most reputation systems on the Web.

To assess the empirical validity of the models, we evaluate them against some external, independently assessed ground-truth. To do so, we measure the accuracy of the model to determine the reputation of known Wikipedia administrators and vandals. We extend the same experiment to Wikipedia good users and blocked users. Furthermore, we measure the predictive value of the proposed reputation models; we calculate reputation of users in English Wikipedia up to time t , then we analyze users’ behavior since then. In aggregate, the results show that the proposed models perform well as classifiers and predictors. Comparison of the model with other similar related work shows that the estimated reputation values are more consistent with users’ behavior.

The remainder of this paper is as follows: Section 2 provides a brief overview of the relevant literature. Section 3 introduces the three reputation models. Section 4 describes the main results including the validation of the three models. Section 5 provides technical background and describes how the Wikipedia data was collected and mined and how the reputation models were implemented. Section 6 provides a discussion and compares the present work to related work in the literature. Finally, Section 7 draws some conclusions and points to a few directions for future investigation.

2 Related Work and Background

Many online communities have trouble motivating enough users to build an active community. High user participation is the key factor for a successful online community, and that is why good motivating factors are essential [14]. As of December 2009, six of the ten most popular Web sites worldwide simply could not exist without user-contributed content [18]. These sites —Myspace, YouTube, Facebook, eBay, Wikipedia, and Craigslist— look for some incentives to encourage broader participation or the contribution of higher quality content. In order to increase and enhance user-generated content contributions, it is important to understand the factors that lead people to freely share their time and knowledge with others [19, 20].

The positive correlation between content quality and user participation discussed in some work [21, 22]. Some studies also showed that building a good reputation/trust can be a motivating factor that encourages user participation in collaborative systems, as well as an incentive for good behavior [23, 24, 25, 26, 14]. There is an extensive amount of research focused on building trust for online communities through trusted third parties or intermediaries [27, 28]. However, it is not applicable to all online communities where users are equal in their roles and there are no entities that can serve as trusted third parties or intermediaries. Reputation management systems provide a way for building trust through social control without trusted third parties [29].

A reputation management system is an approach to systematically evaluate opinions of online community members on various issues (e.g, products, events, etc.) and their opinions concerning the reputation of other community members [30]. Reputation management systems try to quantify reputation based on metrics to rate their users or products. In this way, users are able to judge other users or products to decide on future transactions. A well-known example of reputation management is eBay’s auction and feedback mechanism. In this system, buyers and sellers can rate each other after each transaction by crude +1 or -1 values so that the overall reputation of a trustee becomes the sum of these ratings over the last six months. Besides assigning these ratings, users can add textual annotations to present their experiences during their transactions [31]. In other distributed environments such as peer-to-peer (P2P) file sharing networks or grid computing, users can rate each other after each transaction (e.g., downloading a file). So far, a considerable amount of research has been focused on the development of trust/reputation models in virtual organizations, social networks and P2P networks [32, 33, 34, 35].

Reputation management systems are difficult to scale when they have limited sources of information. Users do not always give feedback about other users/products. They also prefer not to return negative feedback [13]. To overcome this problem, these systems consider reputation as a transitive property and try to propagate it, in order to have an estimation of unknown users and products. In this way, there is a high risk of propagating biased, or inaccurate ratings. An study on P2P e-commerce communities confirms this issue and shows that reputation models based solely on feedback from other peers in the community is inaccurate and ineffective [29]. To smooth out this problem, a reputation management system can make its judgments based on objective observations rather than using explicit experiences from other users; for example, by tracking behavior of users in the system, or analyzing users’ feedback to products over time. Quite unlike some research lines that are based on subjective observations in wiki systems [14, 36], in this work we aim at quantifying reputation based on objective observations of the users’ actions.

The idea of mining history revisions in Wikipedia for inferring reliability of the content was proposed by Zeng *et al.* [37, 38, 39]. To assess reliability of content, the authors take into account the reputation of the authors regardless of behavior. They categorize users into some groups and assign a static reputation value to each group.

Dynamic reputation estimation to wiki users was discussed in some work[5, 40, 41, 42, 43, 44]. Arazy *et al* [5] propose a sentence-ownership algorithm that calculates an author’s contribution to each wiki page. They categorize contributions to wiki pages into four groups: *add content*, *format content*, *add internal link* and *add external link*. Then, they estimate the extent of contributions each author makes to a wiki page. To asses the accuracy of the proposed algorithm, Arazy *et al* compare the author contribution estimated by the proposed algorithm against human judgement for 9 randomly selected articles from Wikipedia. They analyze correlation between the top contributors extracted by their algorithm, and those identified by the human judgement. In summary, the results show that quality of the an author’s contribution to a page is highly correlated with: (a) the number of internal links that the author have added to the page, and (b) the number of sentences contributed by the author that have survived up to the most recent revision of the page.

The most similar work to this paper was presented in [41], where authors assign dynamic reputation to users based on their actions. The estimation process is based on the survival of the text, and survival of the edits. In the system, authors gain reputation when their edits are preserved and they lose reputation when their edits are reverted or undone. Because of the vulnerability of the model to some attacks like *delete-restore* and *fake-followers*, the reputation estimate algorithm is extended in [45] to prevent such

attacks. In this work, we present a robust reputation model that takes into account the survivability of the content. Compared to Adler *et al.* [41], our model is simpler and the reputation estimate is more accurate. In the next section, we explain the process of modeling user reputation in detail.

3 Modeling Reputation

The long-term goal of this effort is to develop an automated system that can estimate the reputation $R_i(t)$ of a Wikipedia user i at time t based on his past behavior. The reputation index $R_i(t)$ should be positive and scaled between 0 and 1 and, for the moment, should be loosely interpretable as the probability that i produces high-quality content. Here we take a first step towards this long term goal by developing several computational models of $R_i(t)$ and testing them, in the form of classifiers, on the available “ground-truth” associated with Wikipedia-known administrators and vandals.

This general approach, which is fairly standard in machine learning applications, requires some explanations. It is reasonable to assume that there exists a true reputation function that is scaled between 0 and 1 and grows monotonically from the user with the lowest reputation to the user with the highest reputation. Our work is an attempt to approximate this unknown function. The only ground-truth available to us concerning this function comes in the form of two extreme datasets of users, the vandals and the admins. No ground-truth data is available for individuals in the middle range of the spectrum. Thus to approximate the true unknown reputation function our first focus is on testing whether the proposed models behave well on the two extreme populations. The models we propose have very few free parameters and they are used to predict reputation values for large numbers of admins and vandals. Once a model capable of producing an output between 0 and 1 for each user has been shown to perform well on the two extreme populations, it is also reasonable to ask whether it performs well on other users. Since no ground truth is available for these users, only indirect evidence can be provided regarding the corresponding performance of the model. Indirect, yet very significant evidence, can be provided in a number of different ways including assessment with respect to other models and data sets proposed in the relevant literature, and results obtained on curated data sets that go beyond the available admin/vandal data. These are precisely the kind of analyses that are described in the following sections.

In order to estimate users’ reputations, we deconstruct edit actions into inserts and deletes. We consider stability of the inserts done by a user, the fraction of inserts that remain, to be an estimate for his reputation. Although stability of deletes can also be considered as another source of information, it has several shortcomings. In fact, Wikipedia is more derived by inserts, and the size of inserts is 1.6 times larger than the size of deletes. Deletes are more difficult to track and therefore calculating stability of deletes is noisier and more computationally extensive. Hence, we make an assumption that using only stability of inserts would result in a reliable estimation of users’ reputation values.

Consider a user i who at time t inserts $c_i(t)$ tokens into a Wikipedia page. It is reasonable to assume that the update $R_i^+(t)$ of $R_i(t)$ should depend on the quality of the tokens inserted at time t . To assess the quality of each token, let t' represent the first time point, after t , where an administrator (hereafter referred to as “admin”) checks current status of a wiki page by submitting a new revision. According to English Wikipedia history dumps, admins on average submit about 11% of the revisions of page, which are distributed over the life cycle of the page.

By definition (or approximation), a token inserted at time t is defined to be of good-quality if it is present after the intervention of the admin at time t' , otherwise it is considered to be of poor-quality. Therefore we have $c_i(t) = g_i(t) + p_i(t)$ where $g_i(t)$ (resp. $p_i(t)$) represents the number of good-quality tokens (resp. poor-quality). For user i , we also let $N_i(t)$ be the total number of tokens inserted up to and right before the time t and, similarly, let $n_i(t)$ be the number of good-quality tokens inserted up to and right before the time t . Using a “+” superscript to denote values immediately after the time t , we have $N_i^+(t) = N_i(t) + c_i(t)$ and $n_i^+(t) = n_i(t) + g_i(t)$.

We can now define three different models of reputation.

Model 1:

In the first model, reputation is simply measured by the fraction of good tokens inserted. In this model, we simply have

$$R_i^+(t) = \frac{n_i^+(t)}{N_i^+(t)} = \frac{n_i(t) + g_i(t)}{N_i^-(t) + c_i(t)} \quad (1)$$

Model 2:

While the first model appears reasonable, tokens that are deleted are treated uniformly. In reality, there is some information to be found in the time at which deletions occur. Vandalistic insertions, for instance, tend to be removed very rapidly [46, 47, 48]. According to our study on Wikipedia, 76% of vandalism is reverted in the very next revision.

Insertions that are deleted only after a very long period of time, tend to be deleted because they are outdated rather than poor in quality. Thus in general, we arrive at the hypothesis that the quicker a token is deleted the more likely it is to be of poor-quality. To realize this hypothesis, we propose a variation on Model 1 where delete tokens introduce a penalty in the numerator with an exponential time decay controlled by a single parameter α .

$$R_i^+(t) = \frac{n_i(t) + g_i(t) - \sum_{d=1}^{p_i(t)} e^{-\alpha(t_d-t)}}{N_i^-(t) + c_i(t)} \quad (2)$$

Here t_d represents the time at which the corresponding token was deleted. Since update rate can vary among different wiki pages, we consider the time interval in terms of the number of revisions. We trained $R_i^+(t)$ over different values of α in order to maximize the area under the ROC curve (AUC). The result shows that $\alpha = 0.1$ returns the best result.

Model 3:

this model is a variation of Model 2 where we take into account also the reputation of the deleter, and use his reputation to weigh the corresponding deletion in the form

$$R_i^+(t) = \frac{n_i(t) + g_i(t) - \sum_{d=1}^{p_i(t)} R_{j(t_d)}(t_d) e^{-\alpha(t_d-t)}}{N_i^-(t) + c_i(t)} \quad (3)$$

The idea behind this variation of the model is to value the deletions done by high-reputation users (e.g. admins) and devalue the deletions done by low-reputation users (e.g. vandals). In Model 3, $\alpha = 0.08$ for the maximum (AUC).

For users who start with a delete action we need to know the initial value, $R_i(0)$. If we denote T the final time, experiments show that the fastest convergence from $R_i(t)$ to $R_i(T)$ is obtained using the initial values $R_i(0) = 0.2$ for all anonymous users, and $R_i(0) = 0.45$ for all registered users (data not shown). These initial values are used in the rest of the paper.

Finally it is worth noting that if Model 3 were to perform well on the classification task (vandals vs admins) this would provide further indirect evidence that Model 3 is self-consistent and may perform well on other users too, since the update equation at time $t + 1$ for Model 3 uses the predicted reputation for users other than vandals or admins at time t .

4 Results

In this section we evaluate the reputation models on our dataset extracted from English Wikipedia history, as described in Section 5.

Table 1: AUC Values for the 3 Reputation Models.

	Admins–Vandals	Good Users–Vandals	Admins–Blocked Users	Good Users–Blocked Users
Model 1	0.9751	0.9839	0.9196	0.9220
Model 2	0.9753	0.9769	0.9094	0.9153
Model 3	0.9742	0.9762	0.9073	0.9125

Table 2: Mean and Standard Deviation of Reputation Values for Admins, Good Users, and Blocked Users for the 3 Reputation Models.

	Admins & Good Users	Blocked Users
Model 1	0.5416 (± 0.2407)	0.0926 (± 0.2091)
Model 2	0.7835 (± 0.1698)	0.1884 (± 0.2962)
Model 3	0.8180 (± 0.1514)	0.2216 (± 0.3128)

4.1 Comparison of Models on Ground Truth Data

We first analyze the performance of the reputation models on two major populations: vandals and admins. Vandals are users who have been blocked by the Wikipedia Committee because they performed edits in violation of Wikipedia rules by engaging in vandalism. The "admin" title is conferred to users selected by the Wikipedia Committee due to their helpful, long-term contributions. Although Model 1, Model 2, and Model 3 have at most one free parameter (α) and can be applied directly to estimate the reputation $R_i(t)$ of any user at any time, here we first use the output $R_i(T)$ to derive a classifier to separate vandals and admins. Table 1 shows the AUC (Area Under the Curve) values corresponding to the ROC curves of the three corresponding classifiers. The table shows that all the three models perform well and their classification performances are comparable. To further analyze classification performance on a broader set of users, we extend the test populations beyond the extreme of vandals and admins to all blocked users on one side and to good users of Wikipedia on the other side.

All blocked users are a superset of the vandals. According to Wikipedia in addition to vandalism, user blocking can happen because of other reasons such as sock-puppetry², edit war³, advertising, or edit disruption. At the other end of the spectrum, automatic extraction of good users beyond admins is not a trivial task. To identify a set of good users, we focus on Wikipedia articles which are marked as good or featured by a committee of experts. From the pool of users contributing to these articles, we extract those who still have contributions that are live in the most recent revisions of these articles. Our definition for good users is also consistent with the result of a recent study of Wikipedia [5], which shows that identification of top page contributors is most highly correlated with the count of their contributed sentences that have survived up to the most recent revision of the wiki pages.

Table 1 shows the AUC values for this extended classification experiment. Similar to the previous results, all the three models perform well and their classification performance are comparable; however, looking at TPRs (True Positive Rates) and FPRs (False Positive Rates) separately (Figure 1) reveals some subtle differences. In particular, we can see that Model 1 is the best model for detecting vandals/blocked users (lower FPR) while Model 3 is the best model for detecting admin/good users (higher TPR).

Table 2 compares the mean and standard deviation of the reputation values for good users and admins against blocked users. In general, all three models assign high reputation values to admins/good users and low reputation values to blocked users; but the distribution of assigned reputations (Figure 2) confirms that Model 1 outperforms the other two models at detecting blocked users, while Model 3 outperforms the other two models at detecting good users.

²http://en.wikipedia.org/wiki/Wikipedia:Sock_puppetry

³http://en.wikipedia.org/wiki/Edit_warring

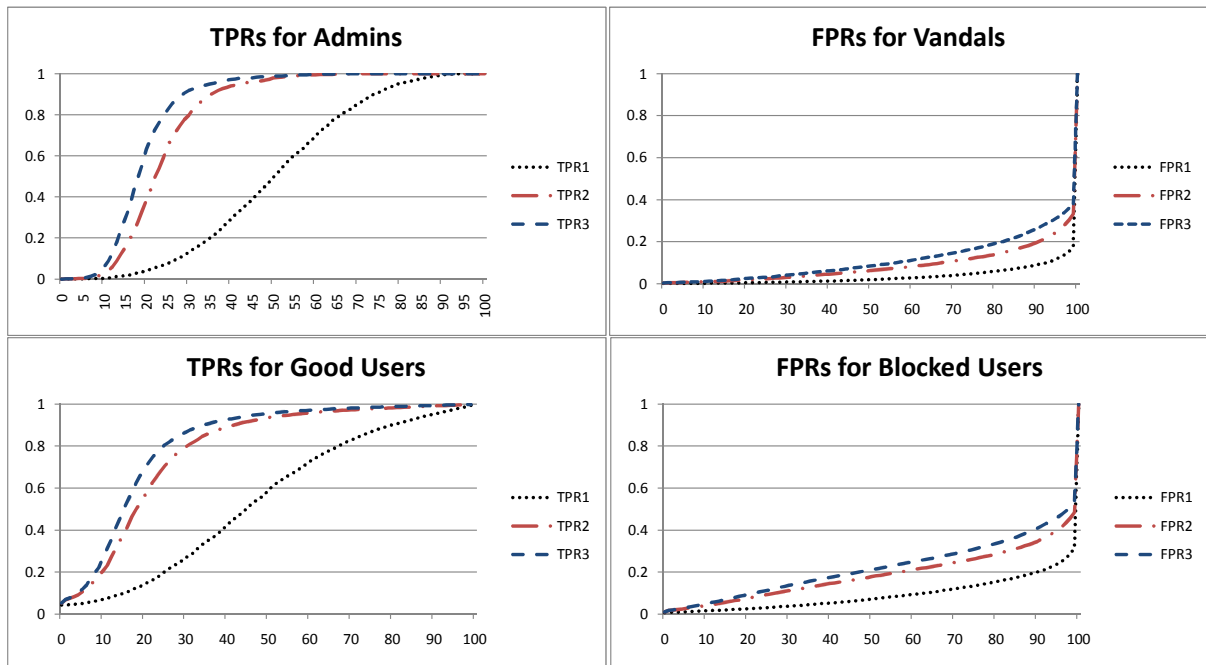


Figure 1: TPRs and FPRs for the 3 Reputation Models as the Classification Threshold is Decreased From 1 to 0.

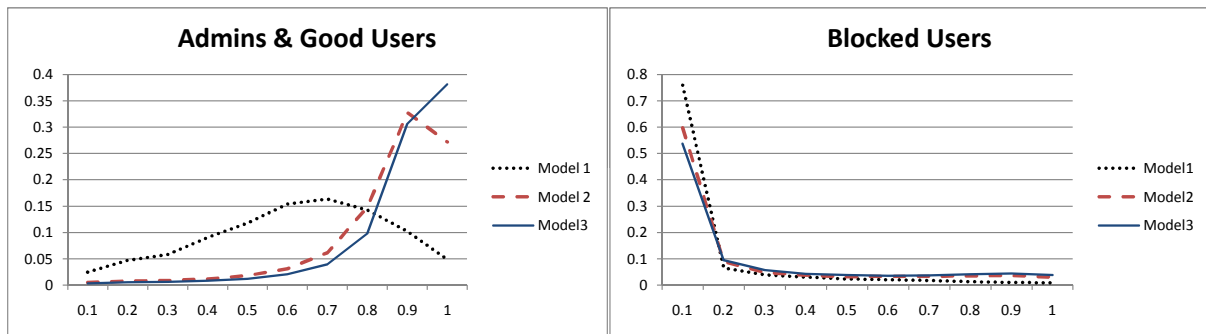


Figure 2: Distribution of Reputation for Good Users/Admins vs Blocked Users Based on the 3 Models.

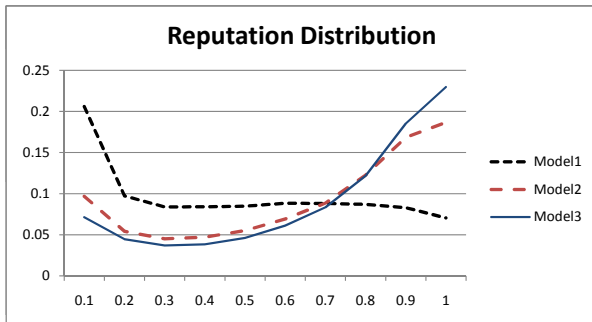


Figure 3: Distribution of Reputation for all Users in English Wikipedia based on the 3 Models.

4.2 Reputation and User Behavior

In this section, we consider the application of the three models to estimate the reputation of all users by extending the previous analyses. We first estimate reputation values for all the users of English Wikipedia. Figure 3 shows the distribution of reputation values for the three models. Unlike Models 2 and 3, where higher reputation users are more dominant, Model 1 yields a higher number of low reputation users. This is a direct consequence of the prompt punishment of a user in Model 1 after his contributed data is deleted. The decrease in reputation punishment occurs in Model 1 regardless of the reason for the deletion or the reputation of the deleter. Hence, it is very likely that Model 1 overly shifts good users to the left. This is also confirmed by the results of the previous experiments and the poor TPRs of Model 1, compared to Model 2 and Model 3.

In order to evaluate the predictive value of the proposed reputation models, we run another experiment. In this experiment, we calculate the reputation of all the users of English Wikipedia up to time t , and analyze the users' behavior up to time t . Then, in a second phase, we analyze their behavior after time t , and correlate this behavior with the reputation values calculated before time t . Specifically we measure the statistical correlation between the reputation of the users at time t and their behavioral indicators before and after time t . We process history revisions up to January 1, 2007 for reputation estimation, and then examine users' behavioral indicators on January 1, 2007 and September 30, 2009.

For each model, we classify all the users into 10 different bins (ignoring bots) according to their reputation values. For each bin associated with model, we calculate the mean of four individual, time-dependent, behavioral indicators, namely RDR, DSR, SDR, and CDR, defined as follows:

- RDR (Reverted Data Ratio) is the ratio of the number of submitted revisions by a user that are reverted by other users, to the total number of revisions submitted by the same user. This metric can be interpreted as the tendency of a user towards contributing vandalistic/problematic content.
- DSR (Data Stability Ratio) is the percentage of contributed data by a user that remains live in the wiki pages. It shows the percentage of content contributed by a user which has not been deleted by other users yet.
- SDR (Submission Data Ratio) is the number of revisions submitted by a user to the total number of submitted revisions. This metric shows how actively each user contributes to the wiki pages by submitting new revisions.
- CDR (Correction Data Ratio) is the ration of the number of reverts done by a user to the total number of reverts. This metric can be interpreted as the tendency of a user to make corrections in the wiki pages.

Figure 4 shows the mean of CDR, SDR, DSR and RDR respectively, in each bin associated with each reputation model, when the behavioral indicators and reputation values are calculated using data up to

Table 3: Correlation Values for the 3 Reputation Models.

Models	RDR	CDR	SDR	DSR
Model 1	(-0.906, -0.871)	(0.434, 0.760)	(0.757, 0.861)	(0.999, 0.996)
Model 2	(-0.927, -0.939)	(0.783, 0.852)	(0.822, 0.833)	(0.976, 0.975)
Model 3	(-0.958, -0.973)	(0.779, 0.811)	(0.791, 0.786)	(0.944, 0.944)

January 1, 2007. As the diagrams show, in general, there is a positive correlation between user reputation and CDR – signifying that users with estimated high reputation tend to make more corrections than users with estimated low reputation. The positive correlation between user reputation and SDR also shows that higher reputation users submit more revisions compared to lower reputation users. The correlation between user reputation and RDR is negative, indicating that lower reputation users tend to contribute vandalistic or low quality content more frequently. These positive and negative correlations are consistent with the general intuitions about Wikipedia that were used to build the models.

It is important to note that among these parameters DSR is a direct input to Model 1 and an indirect input to Models 2 and 3. Hence, the positive correlation between DSR and the user reputation is expected for the three models. For this first set of graphs shown in Fig. 4 this positive correlation does not give any evidence about the predictive value of the models, since both user behavior indicators and user reputation are calculated on the same data.

To show the predictive value of the models, we plot users’ behavioral indicators computed using data up to September 30, 2009 against the reputation values estimated using data up to January 1, 2007. Figure 5 shows the mean of CDR, SDR, DSR and RDR respectively, in each bin associated with each reputation model, where the users’ behavioral indicators are estimated in 2009, while reputation values used to determine the bins are estimated at the beginning of 2007. The first observation is that this second set of curves has similar shapes to those in Figure 4, indicating that the estimated users’ reputations are consistent with their behaviors– users continue to behave in 2007–2009 as they had behaved before 2007. Furthermore, behavior or reputation is captured in broad strokes, by the reputation models.

The values of the behavioral indicators in Figure 5 are slightly different from their predicted values corresponding to Figure 4. For example, according to Model 3 applied up to 2007, users with a reputation of 0.1 or below ought to have 69% reverted data (RDR), whereas in reality during 2007–2009 those users had only 52% reverted data. Likewise, the same Model 3 predicts that users with a reputation between 0.8 and 0.9 ought to be responsible for 37% of the total number of submissions (SDR), whereas in reality during 2007–2009 those users were responsible for only 27% of submissions. To compare these two sets of diagrams (Figure 4 and Figure 5), we perform a Pearson correlation analysis. The results are described in Table 3, where each tuple shows the correlation between the two parameters before and after 2007, respectively. For example, the entry (-0.906, -0.871) signifies that the correlation between RDR and Model 1 reputation is -0.906 in Figure 4, while it is -0.871 in Figure 5. These correlations are highly significant and the same is observed in one measures the correlation between the reputation value themselves within or across models, and up to 2007 or up to 2009.

In combination, these results suggest that the reputation models are good at predicting behavioral indices and reputation values at future times, not only for extreme populations of very good of very bad users, but across the entire spectrum of reputation values.

5 Tools and Methods

In order to get the data for our study, we used 5 client machines for a period of 2.5 months during summer 2009 to send requests to MediaWiki API and extract the data. By sending consecutive requests to MediaWiki API, one can get the text of all revisions of each Wikipedia article. We needed the list of the articles in English Wikipedia to feed to the API in order to get article revisions. However, a

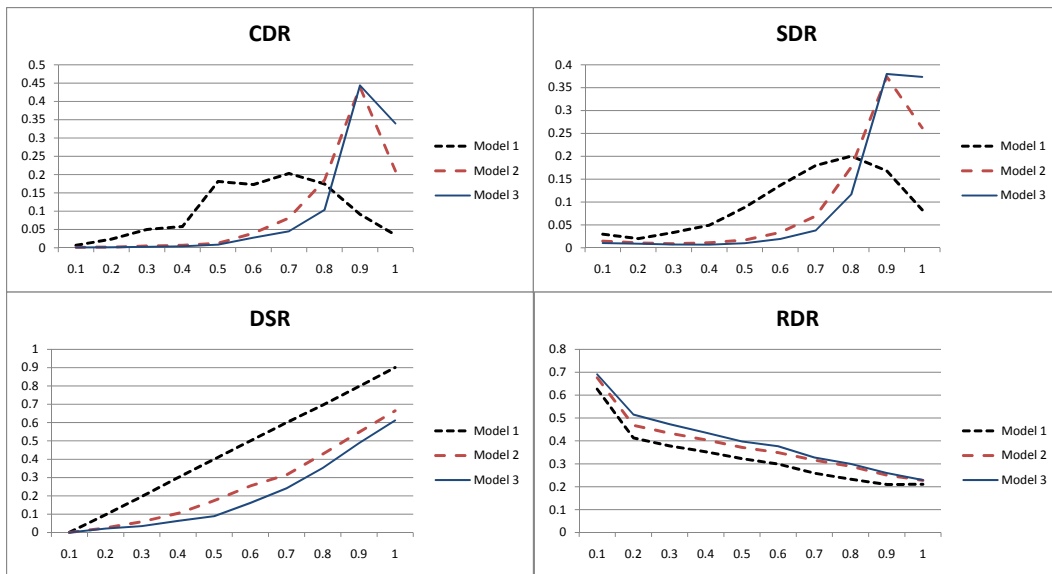


Figure 4: CDR, SDR, DSR and RDR as Functions of Reputation (Base on Data Before 2007).

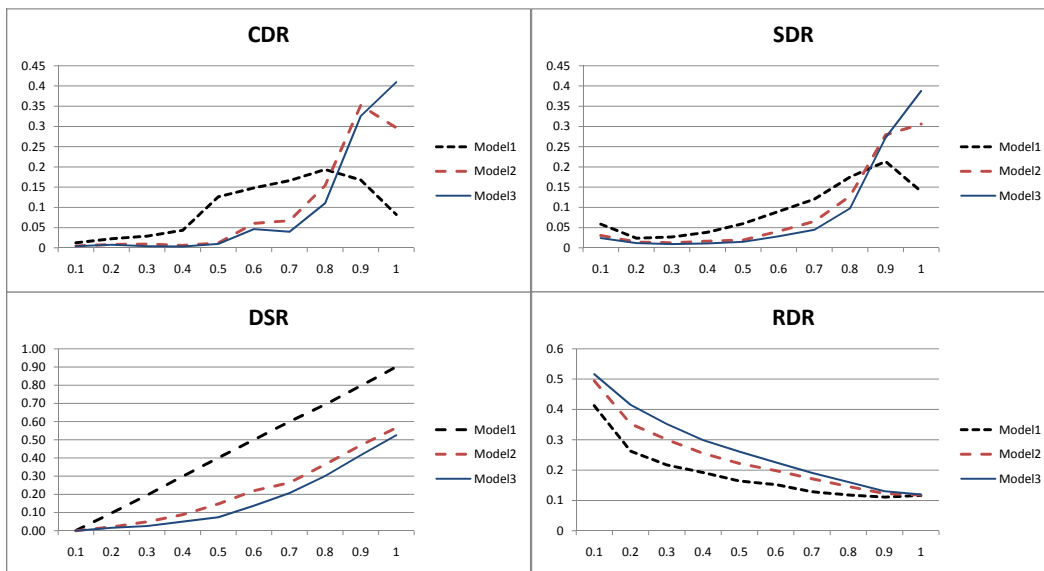


Figure 5: CDR, SDR, DSR and RDR Extracted After 2007 as Functions of Reputation Computed Before 2007.

Table 4: Properties of the Data Set

Time Span	96 months
Number of Users	12, 797, 391
Registered Users	1, 749, 146
Anonymous Users	11, 048, 245
Number of Articles	1, 899, 622
Featured	2, 650
Good Users	197, 436
Good	7, 502
Good Users	334, 369
For Deletion	125
Regular	1, 889, 345
Number of Revisions	123, 938, 034
by Anonymous Users	82, 577, 828
by Registered Users	41, 360, 206

significant number of Wikipedia articles had been redirected to other articles which we ignored them. In order to obtain a clean list of Wikipedia articles, we used crawler4j [49] to crawl English Wikipedia and extract the list of non-redirected articles. We started from the Wikipedia main page and some other seed pages and by traversing the links we crawled about 1.9 million articles. We also used the MediaWiki API to extract different types of contributors such as *bots*⁴, *admins* and blocked users. Table 4 shows the properties of the data set.

A note about “users”. It is virtually impossible to associate actual persons with the internet behavior in a one to one fashion. To bypass this problem Wikipedia defines two classes of users. An anonymous user is a user that is known only through his IP address. A registered user is a user associated with his usernames (i.e. nicknames) that was entered during the registration process. We, as well as others [50, 51, 52], follow the same nomenclature as Wikipedia: a user in this study refers to a registered account or an IP address, and it does not refer to a real-world individual.

5.1 Extracting Reverts

A revert is an action to undo all changes made to an article and is primarily used for fighting vandalism. For extracting reverts, we compare the text of each revision to the text of the previous revisions. Since the text comparison process is computationally expensive, the comparison is done on the MD5 signature of the texts rather than on the texts themselves.

5.2 Extracting Events

We consider an atomic event to be an insertion or deletion of a word. Insertions are extracted by comparing the text of each revision with the text of the previous revision; deletions are extracted by comparing the text in a revision with the text of the all subsequent revisions. We use the diff algorithm described in [53], for accurate extraction of atomic events. The advantage of this algorithm compared to most of current diff algorithms is its ability to detect movements of blocks. The developed tool, named

⁴Bots are generally programs or scripts that make automated edits without the necessity of human decision-making.

Wikipedia Event Extractor, is publicly available at [54]. We calculated $R_i(T)$ of users by processing the extracted events.

6 Discussion and Comparison to Related Work

In this section, we discuss our model in more detail and compare it to related work in the literature according to several different criteria, appearing in boldface in the list below. criteria:

Tracking Token Ownership.

Effective assignment of inserts and deletes to owners is highly dependent on: (1) the accuracy of the diff algorithm used for calculating the distance between two revisions of a wiki page; (2) the side-effects of reverts resulting in incorrect ownership assignments. An effective diff algorithm for wikis should identify differences in a way that is meaningful to human readers. In particular, reordering of text blocks should be detected in order to accurately assign ownership to the tokens in the reordered blocks. This issue has not been taken into consideration in some of the previous work [36, 55, 5]. For example Sabel *et al.* [36] use the Levenshtein algorithm⁵ to compute the edit distance between two revisions. This algorithm penalizes block reordering and as a result each token that has been shifted is usually considered deleted from its old position, and inserted in its new position [56, 57]. In our experience, the Wikipedia’s diff algorithm can suffer from the same problem, occasionally preventing the detection of block reorderings. We and others [41] overcome this problem by using efficient diff algorithms that detect reordering of blocks and run in time and space linear to the size of the input [53, 58].

Another issue in accurate assignment of token ownership has to do with taking into account the side-effects of reverts. In general, successive revisions of a wiki page have similar content, and each revision, except the very first, is a descendant of the preceding one. However, this model is insufficient for describing the realistic evolution of a wiki page [36]. Assume that a vandal blanks out the i th revision of a wiki page. Therefore, the $(i + 1)$ th revision becomes blank. When user u reverts the $(i + 1)$ th revision to the previous revision, this revert results in a new revision and the content of $(i + 2)$ th revision and i th revision become the same. This scenario raises several problems: (1) users whose contributions were deleted by the vandal are penalized unfairly; (2) u is erroneously considered to be the owner of all the content of the $(i + 2)$ th revision; (3) the true original owner(s) are denied ownership of the content they actually contributed. We and others [41] address this issue by ignoring these spurious insertions and deletions caused by reverts. However, in [41], the authors decided to process only up to the 3rd successive revision in order to extract reverts and assign ownership. Our study of Wikipedia shows that about 6% of reverts return the i th revision of a page to the j th where $i - j > 3$. For this reason, in order not to lose any information, we process all revisions. Because reverts happen very frequently in Wikipedia, ignoring the side-effect of reverts can result in significant numbers of incorrect assignments of token ownership.

Stability of Edits.

For the purpose of this study, user reputation is estimated by looking at the stability of the content he contributes. To estimate the stability of the content, we track the tokens inserted by a user up to the last revision of the page to see how many of these tokens are deleted. In some of the related work in the literature, the tracking process has been more limited, for instance by tracking inserted tokens only up to a limited number of successive revisions and therefore missing some deleted tokens. For example, the authors in [41] use up to the 10th successive revisions. Our study of Wikipedia shows that 37% of the deletes happen after the 10th revision. Hence, ignoring this fraction of deletes may lead to reputation estimates that are less accurate. For the purpose of this study user reputation is estimated by considering the stability of inserts only. One may argue that although the number of deletes is considerably smaller

⁵http://en.wikipedia.org/wiki/Levenshtein_distance

than the number of inserts, there is some information in the stability of the deletes too, and one ought to be able to use this additional information to derive even more accurate models of reputation. To see if the stability of deletes can improve the accuracy of the models, we reformulate our simplest model (Model 1) by considering the stability of deletes. We define Model 1' as follows,

$$R_i^+(t) = \frac{n_i^+(t) + n_d^+(t)}{N_i^+(t) + N_d^+(t)} \quad (4)$$

where $n_d(t)$ is the number of good-quality deleted tokens and $N_d^+(t)$ is the total number of deleted tokens after time t . We tested Model 1' as a classifier on admins and vandals and the results showed that Model 1' has lower AUC (0.84) than Model 1. Interestingly this observation is consistent with the result of another study [5], which shows that delete and proofread edits have little impact on the perception of top contributors in Wikipedia. In other words, there does not seem to exist any significant correlation between an author's reputation and an author's number of deletes in the wiki pages; but, in contrast, there is a very strong correlations between an author's reputation and an author's number of insertions.

Dynamic/Non-Dynamic and Individualized/Non-Individualized Reputation Measures.

One of the advantages of the models presented here is that they assign individualized and dynamic reputation values to both anonymous and registered users. This is not the case in some of the related work published in the literature. For example, the authors in [37], use non-dynamic and non-individualized reputation values for the users. They categorize users into four groups —administrators, anonymous users, registered users, and blocked users— and assign a static reputation value to each group. In [41], authors consider dynamic and individualized reputation values for registered users, but assign a static and non-individualized reputation value to anonymous users.

Resistance to Attacks.

According to the proposed models, users increase their reputation when their contributions to the wiki pages survive. The robustness of the models are highly dependent on when the reputation gain events are triggered. Assume that the reputation of a user increases immediately after he inserts some content; if the page is revised only after a long period of time, the user will have an increased reputation throughout the period, even if his contribution is of poor quality. One solution to this problem is to postpone the reputation increase until the contribution is reviewed by another user. Although this solution solves the previous problem, the reputation model becomes vulnerable to a Sybil attack⁶, in which an attacker has multiple identities and can follow up his own edits. To overcome both problems at once, we postpone the reputation increase until a high-reputation user (*e.g.* admin) approves the corresponding page. Therefore, in the proposed models, a reputation gain can be triggered only when an admin submits a new revision. One may argue that this reliance on the limited number of admins as outside authorities might reduce the accuracy or scope of applicability of the proposed models. However, as shown in Table4, in Wikipedia we have large number of good users which contribute actively to Wikipedia pages. Thus enlarging the pool of authorities beyond admins to include these good users to validate the quality of insertions may provide an efficient solution, especially for pages with high edit rates.

Among related work, Adler *et al.* [45] has addressed the attack resistance problem by extending their previously presented model [41]. Although the extended model is resistant to the aforementioned attacks, it is considerably more complex than the original model. Since we do not consider the stability of deletes and reverts and we ignore the side-effect of reverts, our proposed models are not prone to other kinds of attacks, such as delete-restore or fake-followers [45].

Another issue in the proposed models is that reputation gains happen without giving any consideration to the quality of the page that a user contributes to. In [42], the authors make two assumptions: (1) the quality of a wiki page depends on the reputation of its contributors; (2) the reputation of a user depends on the quality of the pages he contributes to. Although the first assumption is often true, the second

⁶http://en.wikipedia.org/wiki/Sybil_attack

assumption is more debatable; furthermore, it also increases the vulnerability of the model against some attacks. Our study of Wikipedia shows that vandals are more active in high-quality pages. For example, the average RDR (Reverted Data Ratio) associated with featured articles⁷ is 17.8% (11.4% before being marked as featured and 25.4% after being marked as featured) while it is about 9.9% for non-featured articles. In general, a policy based on the assumptions in [42], would result in vandals having more incentives to contribute to high-quality pages hoping to increase their reputations, and high reputation users having less incentives to contribute to low-quality pages to improve their quality.

Population Coverage and Precision and Recall Issues.

In related work, anonymous users are either completely ignored or assigned a *static* reputation value, regardless of their behavior [41]. There are three main reasons why we think that it is important to consider anonymous users in the reputation estimation process: (1) About 33% of the submissions and 39% of the inserts in Wikipedia are contributed by anonymous users and 16% of these contributions have survived up to the last revisions of the articles, therefore they cannot be ignored; (2) Wikipedia itself blocks IP addresses associated with anonymous vandals and 40% of anonymous vandals are subject to infinite blocking. Therefore, an effective reputation management system for Wikipedia should be able to identify anonymous vandals; otherwise, a significant number of vandals will go undetected; and (3) About 15% of data deleted from registered users is deleted by anonymous users, hence ignoring their deletes would degrade the accuracy of the estimated reputation for registered users.

To further verify the relevance of anonymous users, we reformulate Model 3 and assign a static reputation value to all anonymous users, as suggested in [41, 37]. Several static reputation values were tested and the results for the new model (Model 3') show that the AUC always drops, for instance by 1% when the reputation of all anonymous users is set to 0.1. These results indicate that ignoring the anonymous population is likely to decrease the accuracy of a reputation model.

Evaluation results reported by Adler *et al* [41] using a precision and recall analysis also confirm this observation. To be more specific, in their work they use a model to estimate reputation values up to time t , and then estimate the precision and recall after time t provided by *low reputation users* for *short-lived text*, which are defined as follows:

- Short-lived text is text that is almost immediately removed (only 20% of the text in a version survives to the next version).
- A low-reputation author is an author whose reputation falls in the bottom 20% of the reputation scale.

Table 5 shows the precision and recall values obtained on these data by Adler *et al* by first ignoring anonymous users (first row) and then by assigning a static common reputation value to all anonymous users (second row). The third row shows the results obtained using Model 3, the most similar of our models to their model, to estimate reputations in English Wikipedia up to 2007 and measure precision and recall on the same data. As the table shows, the model by Adler *et al* [41] performs better when a reputation is assigned to anonymous users, albeit statically. Model 3 significantly outperforms the other two approaches because of dynamic assignment of reputation to anonymous users, better token ownership assignments, and also effective removal of side effects of reverts.

7 Conclusions and Future Work

In this paper, we have modeled user reputation in wiki systems. We have presented 3 reputation models. According to Model 1, when a user inserts a piece of content in a wiki page his reputation is increased and when a piece of content contributed by the user is deleted by another user, his reputation is decreased. In Model 2, we also take into account the time interval between insertions and deletions of content items. Finally, in Model 3, we add another parameter which is the current reputation of the deleter.

⁷http://en.wikipedia.org/wiki/Featured_Article

Table 5: Precision and Recall Provided by Low Reputation Users for Short-Lived Text.

Models	Precision	Recall
Ignoring Anonymous Users [41]	0.058	0.378
Considering Anonymous Users [41]	0.190	0.904
Model 3	0.404	0.975

We have tested the three models on English Wikipedia which is the largest online wiki with an open editing model, allowing anyone to enter and edit content. Our experiments show that the three models can accurately assign reputation values to Wikipedia’s known administrators/good users and vandals/blocked users. Additional analyses reveal that Model 1 does slightly better at detecting vandals and Model 3 does slightly better at detecting good users.

The proposed models can be applied in real time to calculate dynamic and individualized reputation values. While Model 1 is simpler to implement, Model 3 appears to be slightly more accurate, and more robust to attacks than several other competing models of reputation. We are currently exploring several directions for improving the proposed models, for instance, by combining several models (including RDR and DCR) in different ways, and incorporating more information about deletion behavior.

Since all wikis store history pages, the proposed models can be used in any intra-company or intra-organization wikis or in public wikis such as Wikipedia, Citizendium⁸, and Scholarpedia⁹ and they can be integrated with the corresponding platform. In addition, most of the online wikis like Citizendium and Scholarpedia use the same wiki software, MediaWiki, as Wikipedia. Therefore, the implementation of the models can be modified easily to fit those other platforms. The proposed models can be used in wikis for rating users or as a decision support system for administrators. For example, they can be used for automatic vandal detection, saving substantial amounts of time for wiki administrators. They can be integrated also in a quality assessment system that assesses the reliability of the content according to the reputation of its contributors, as suggested in [42].

In future work, we would like to develop a search engine for Wikipedia, capable of returning the most recent and reliable revision of each wiki page when searching for a topic in Wikipedia¹⁰. For each page, the most recent reliable revision can be defined as the most recent revision submitted by a user with a high reputation.

8 Acknowledgment

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

References

- [1] T. O’Reilly, “What is web 2.0: Design patterns and business models for the next generation of software,” *Communications & Strategies*, 2007.
- [2] B. Alexander, “Web 2.0: A new wave of innovation for teaching and learning?” *Educause Review*, vol. 41, no. 2, pp. 32–44, 2006.
- [3] M. Boulos, I. Maramba, and S. Wheeler, “Wikis, blogs and podcasts: a new generation of web-based tools for virtual collaborative clinical practice and education,” *BMC Medical Education*, vol. 6, p. 41, 2006.

⁸<http://en.citizendium.org/>

⁹<http://www.scholarpedia.org/>

¹⁰<http://wikijoo.org/>

- [4] B. Leuf and W. Cunningham, *The Wiki Way: quick collaboration on the web*. Addison-Wesley, 2001.
- [5] O. Arazy and E. Stroulia, “A utility for estimating the relative contributions of wiki authors,” in *International AAAI Conference on Weblogs and Social Media*, 2009. [Online]. Available: <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/157>
- [6] R. Raitman and N. Augar, “Employing wikis for online collaboration in the e-learning environment: Case study,” in *Proceedings of the Third International Conference on Information Technology and Applications (ICITA’05)*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 142–146.
- [7] E. Elrufaie and D. A. Turner, “A wiki paradigm for use in it courses,” in *ITCC ’05: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC’05) - Volume II*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 770–771.
- [8] Q. Li, “Knowledge building community: Keys for using online forums prediction,” *TechTrends*, pp. 24–29, 2007.
- [9] J. Giles, “Internet encyclopaedias go head to head,” *Nature*, pp. 438:900–901, December 2005.
- [10] J. Seigenthaler. (2005) A false wikipedia ‘biography’. [Online]. Available: http://www.usatoday.com/news /opinion/editorials/2005-11-29-wikipedia-edit_x.htm
- [11] B. Shneiderman, “Designing trust into online experiences,” *Communications of ACM*, vol. 43, no. 12, pp. 57–59, 2000.
- [12] P. Resnick and R. Zeckhauser, “Trust among strangers in Internet transactions: Empirical analysis of eBay’s reputation system,” in *The Economics of the Internet and E-Commerce*, ser. Advances in Applied Microeconomics, M. R. Baye, Ed. Elsevier Science, 2002, vol. 11.
- [13] A. Josan, R. Ismail, and C. Boyd, “A survey of trust and reputation systems for online service provision,” *Decis. Support Syst.*, vol. 43, no. 2, pp. 618–644, 2007.
- [14] B. Hoisl, W. Aigner, and S. Miksch, “Social rewarding in wiki systems motivating the community,” in *Proceedings of HCI International - 12th International Conference on Human-Computer Interaction (HCII 2007)*, vol. 4564/2007. LNCS, Springer, 2007, pp. 362–371.
- [15] Reputation. [Online]. Available: <http://en.wikipedia.org/wiki/Reputation>
- [16] C. L. Corritore, B. Kracher, and S. Wiedenbeck, “On-line trust: concepts, evolving themes, a model,” *International Journal Human-Computer Studies*, vol. 58, no. 6, pp. 737–758, 2003.
- [17] Knowledge smackdown: Wikipedia vs. citizendium. [Online]. Available: http://www.storysouth.com/comment /2006/09/knowledge_smackdown_wikipedia.html
- [18] Alexa’s top 10 websites. [Online]. Available: <http://www.alexa.com/>
- [19] O. Nov, “What motivates wikipedians?” *Commun. ACM*, vol. 50, no. 11, pp. 60–64, 2007.
- [20] A. H. Maslow, *Motivation and personality*. HarperCollins Publishers, 1987.
- [21] D. M. Wilkinson and B. A. Huberman, “Cooperation and quality in wikipedia,” in *WikiSym ’07: Proceedings of the 2007 international symposium on Wikis*. New York, NY, USA: ACM, 2007, pp. 157–164.
- [22] Y. Ganjisaffar, S. Javanmardi, and C. Lopes, “Review-based ranking of wikipedia articles,” in *Proceedings of the International Conference on Computational Aspects of Social Networks*, June 2009.

- [23] D. Anthony, S. W. Smith, and T. Williamson, “Explaining quality in internet collective goods: zealots and good samaritans in the case of wikipedia,” Hanover : Dartmouth College, Tech. Rep., 2005. [Online]. Available: web.mit.edu/iandeseminar/Papers/Fall2005/anthony.pdf
- [24] J. Voss, “Measuring wikipedia,” in *Proceedings of 10th International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden, 2005.
- [25] F. B. Viegas, 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*. New York, NY, USA: ACM, 2004, pp. 575–582.
- [26] A. Josang, C. Keser, and T. Dimitrakos, “Can we manage trust?” in *Proceedings of the Third International Conference on Trust Management (iTrust'05)*, May 2005.
- [27] S. Ketchpel and H. Garcia-molina, “Making trust explicit in distributed commerce transactions,” in *In Proceedings of the International Conference on Distributed Computing Systems*, 1996, pp. 270–281.
- [28] Y. Atif, “Building trust in e-commerce,” *IEEE Internet Computing*, vol. 6, no. 1, pp. 18–24, 2002.
- [29] L. Xiong and L. Liu, “A reputation-based trust model for peer-to-peer ecommerce communities [extended abstract],” in *EC '03: Proceedings of the 4th ACM conference on Electronic commerce*. ACM, 2003, pp. 228–229.
- [30] A. Gutscher, “A trust model for an open, decentralized reputation system,” *Trust Management*, pp. 285–300, 2007.
- [31] (2005) Evaluating a member’s reputation. [Online]. Available: <http://pages.ebay.com/help/feedback/evaluating-feedback.html>
- [32] A. C. Squicciarini, F. Paci, and E. Bertino, “Trust establishment in the formation of virtual organizations,” in *ICDE Workshops*, 2008, pp. 454–461.
- [33] R. Aringhieri, E. Damiani, S. D. C. D. Vimercati, S. Paraboschi, and P. Samarati, “Fuzzy techniques for trust and reputation management in anonymous peer-to-peer systems: Special topic section on soft approaches to information retrieval and information access on the web,” *Journal of the American Society for Information Science and Technology*, vol. 57, no. 4, pp. 528–537, 2006.
- [34] C.-N. Ziegler and J. Golbeck, “Investigating interactions of trust and interest similarity,” *Decis. Support Syst.*, vol. 43, no. 2, pp. 460–475, 2007.
- [35] H. Liu, E.-P. Lim, H. W. Lauw, M.-T. 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*. New York, NY, USA: ACM, 2008, pp. 310–319.
- [36] M. Sabel, A. garg, and R. Battiti, “Wikirep: Digital reputation in virtual communities,” University of Trento, Tech. Rep., 2005. [Online]. Available: <http://eprints.biblio.unitn.it/archive/00000810/>
- [37] 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.
- [38] H. Zeng, M. Alhossaini, R. Fikes, and D. McGuinness, “Mining revision history to assess trustworthiness of article fragments,” in *In Proceedings of the 2nd International Conference on Collaborative Computing: Networking, Applications, and Worksharing*, 2006.

- [39] D. L. McGuinness, H. Zeng, P. P. da Silva, L. Ding, D. Narayanan, and M. Bhaowal, “Investigations into trust for collaborative information repositories: A wikipedia case study,” in *Proceedings of the Workshop on Models of Trust for the Web*, May 2006.
- [40] S. Javanmardi and C. Lopes, “Modeling trust in collaborative information systems,” in *Collaborate-Com '07: Proceedings of the 3rd International Conference on Collaborative computing: Networking, Applications and Worksharing*. New York, NY, USA: IEEE, November 2007.
- [41] 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.
- [42] 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*. ACM, 2007, pp. 243–252.
- [43] M. Sabel, “Structuring wiki revision history,” in *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*. ACM, 2007, pp. 125–130.
- [44] N. T. Korfiatis, M. Poulos, and G. Bokos, “Evaluating authoritative sources in collaborative editing environments,” *Online Information Review*, vol. 30, no. 3, pp. 252–262, 2006.
- [45] K. Chatterjee, L. de Alfaro, and I. Pye, “Robust content-driven reputation,” School of Engineering, University of California, Santa Cruz, CA, USA, Tech. Rep. UCSC-SOE-08-09, 2008. [Online]. Available: <http://www.soe.ucsc.edu/luca/papers/08/ucsc-soe-08-09.pdf>
- [46] F. B. 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.
- [47] 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.
- [48] P. D. Magnus, “Early response to false claims in wikipedia,” *First Monday*, vol. 13, no. 9, September 2008.
- [49] Crawler4j. [Online]. Available: <http://crawler4j.googlecode.com/>
- [50] 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.
- [51] 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.
- [52] M. Ekstrand and J. Riedl, “rv you’re dumb: Identifying discarded work in wiki article history,” in *WikiSym '09 Proceedings of the 2009 International Symposium on Wikis*. ACM, October 2009.
- [53] P. Heckel, “A technique for isolating differences between files,” *System Sciences*, pp. 264–268, 1978.
- [54] Wikipedia event extractor. [Online]. Available: <http://mondego.calit2.uci.edu/WikipediaEventExtractor/>
- [55] M. Hess, B. Kerrand, and L. Rickards, “Wiki user statistics for regulating behaviour,” Tech. Rep., 2006. [Online]. Available: <http://icd.si.umich.edu/684/files/684%20wikistat%20paper%202.pdf>

- [56] G. Leusch and H. Ney, “Bleusp, invwer, cder: Three improved mt evaluation measures,” in *NIST Metrics for Machine Translation Challenge*, Waikiki, Honolulu, Hawaii, October 2008.
- [57] G. Leusch, N. Ueffing, and H. Ney, “Cder: Efficient mt evaluation using block movements,” in *In Proceedings of EACL*, 2006, pp. 241–248.
- [58] W. F. Tichy, “The string-to-string correction problem with block moves,” *ACM Transactions on Computer Systems*, vol. 2, no. 4, pp. 309–321, 1984.