A Content-Driven Reputation System for the Wikipedia

B. Thomas Adler
Computer Science Dept.
University of California
1156 High St., Santa Cruz, CA 95064, USA

Luca de Alfaro
Computer Engineering Dept.
University of California
1156 High St., Santa Cruz, CA 95064, USA

November 21, 2006

Technical Report ucsc-crl-06-18
School of Engineering
University of California, Santa Cruz

Abstract

On-line forums for the collaborative creation of bodies of information are a phenomenon of rising importance; the Wikipedia is one of the best-known examples. The open nature of such forums could benefit from a notion of reputation *for its authors. Author reputation could be used to flag new contributions from low-reputation authors, and it could be used to allow only authors with good reputation to contribute to controversial or critical pages. A reputation system for the Wikipedia would also provide an incentive to give high-quality contributions.*

We present in this paper a novel type of content-driven *reputation system for Wikipedia authors. In our system, authors gain reputation when the edits and text additions they perform to Wikipedia articles are long-lived, and they lose reputation when their changes are undone in short order. We have implemented the proposed system, and we have used it to analyze the entire Italian and French Wikipedias, consisting of a total of 691,551 pages and 5,587,523 revisions. Our results show that our notion of reputation has good* predictive *value: changes performed by low-reputation authors have a significantly larger than average probability of having poor quality, and of being undone.*

1 Introduction

The collaborative, web-based creation of bodies of knowledge and information is a phenomenon of rising importance. Arguably the most successful example of collaborative content creation, as well as one of the oldest, is the Wikipedia.[1] The Wikipedia is a set of encyclopedias, each in a different language, that cover a very wide swath of knowledge, from history to geography, from math to politics, from pop culture to archeology. Anyone can contribute to the Wikipedia: when an article is displayed, the reader can click on an "edit" button, and is thus able to modify the content of the article. Subsequent users can improve on the contribution, or revert the article to its previous form. An important feature of the Wikipedia is that for each article, all revisions are kept. This makes it easier to undo edits than to perform them: in particular, it makes it easy for good users to undo the improper edits performed by rogue users. Since good users outnumber rogue ones, a fundamental insight behind wiki development was that good content would predominate [CL01].

In spite of the self-policing character of the Wikipedia, its notoriety has attracted many rogue contributors, who have defaced or inserted false or inappropriate material. While this material is quickly removed, it was felt that this was enough of a problem that many articles dealing with high-visibility or controversial topics have been "protected": only authors who have contributed to the Wikipedia for some time may edit them.[2] The length of time for which a user has contributed to the Wikipedia is, of course, a basic form of reputation. Another criticism leveraged at the Wikipedia is that, since contributions are not signed with the real names of the authors, a user lacks an easy way of assessing the reliability of the text in an article; this problem received much attention in the press (see e.g. [Str06, HR06]). While the average quality of Wikipedia articles is high [Gil05], it may be useful to know which text was inserted by experienced authors,

[1] www.wikipedia.com

[2] http://en.wikipedia.org/wiki/Wikipedia:Semi-protection_policy

and lasted through many edits, and which text has been just inserted by authors who have a poor track record: in other words, it may be useful to know the reputation of the users who contributed, and vetted, the text. In this paper, we present a reputation system for Wikipedia authors that may help fulfill this need. Although we do not advocate, in this paper, a particular application of author reputation over another, we mention below some of the most obvious possibilities:

- **Reputation-based text coloring.** Each article could display a button labeled "check text reputation": upon clicking the button, a user would be led to a copy of the page, where the text background color reflects the reputation of the author of each portion of text, as well as the reputation of authors who vetted the text, editing the page while leaving the text in place [Cro06]. The appeal of this method is that reputation is displayed in an anonymous way, associated to the article text. This avoids placing blame or praise directly on the authors: the impersonal character of this feedback could be well-suited to a collaborative forum such as the Wikipedia.

- **Restricting edits.** Highly controversial articles could be protected, so that only authors with sufficiently high reputation are able to edit them.

- **Alerting editors about low-reputation edits.** Wikipedia Editors keep a watchful eye on most controversial articles, and in fact, on a large portion of the Wikipedia, improving content and undoing poor-quality revisions. A reputation system could be used to alert them whenever a crucial or controversial article is modified by a low-reputation author.

- **Provide an incentive for high-quality contributions.** A reputation system could provide an additional incentive for authors to provide high-quality contributions to the Wikipedia.

1.1 A Content-Driven Reputation System

Most reputation systems are *user-driven:* they are based on users rating each other's contributions or behavior [RZFK00, Del03]. A famous example is the rating system of Ebay, where buyers and sellers rate each other after performing transactions. In contrast, the reputation system we propose requires no user input: rather, Wikipedia authors are evaluated on the basis of how their contribution to the Wikipedia fares. More precisely, suppose that an author A contributes to a Wikipedia article by editing it. When another author B subsequently revises the same article, she may choose to preserve some of the edits performed by A. By preserving them, B provides her vote of confidence in these edits, and in author A. Our reputation system will increase the reputation of A in a manner that depends on the amount of preserved edits, as well as on the reputation of B.

We call such a reputation system, based on content evolution, rather than on user input, a *content-driven* reputation system. A content-driven reputation system ensures that an author's reputation depends only on how the author's contributions fare in the Wikipedia: as such, it has an intrinsic objectivity advantage over user-driven reputation systems. In order to *badmouth* (damage the reputation of) author B, an author A cannot simply give a negative rating to a contribution by B. Rather, to discredit B, A needs to undo some contribution of B, thus running the risk that if subsequent authors restore B's contribution, it will be A's reputation, rather than B's, to suffer. Likewise, authors cannot simply praise each other's contributions to enhance their reputations: their contributions must actually withstand the test of time.

Admittedly, a content-driven reputation system can be less accurate than a user-driven one. Author contributions can be deleted for a variety of reasons, including reorganizations and thorough rewrites of the articles. Our reputation system copes with this in two ways. First, the way we assign reputation is able to distinguish between edits that are later reverted, and edits that are later further refined. The reputations of authors of reverted edits suffer; the reputations of authors of further refined edits do not. Hence, spammers, whose contributions are reverted, suffer, while contributors to initial versions of articles, which will be extensively rewritten, do not, even though both kinds of contributions do not survive in the long term. Second, contributions that are appropriate, but that can be improved, tend to last longer than inappropriate contributions, thus receiving at least partial credit.

While it could be arguably interesting to explore combinations of user-generated and content-driven reputation, in this paper we focus on content-driven reputation alone. The computational nature of content-driven reputation enables us to evaluate its effectiveness directly on the Wikipedia: building a separate (and, inevitably, small-scale and low-traffic) test wiki to experiment with the ideas is not necessary. We will present extensive data on the accuracy and performance of our content-driven reputation system, by applying it to the revision history of the Italian and French Wikipedias. The data will show that the value of content-driven reputation can be used to predict the quality of author's contributions.

1.2 Prescriptive, Descriptive, and Predictive Reputation

Reputation is most commonly used for its *prescriptive* and *descriptive* value:

- *Prescriptive value.* Reputation systems specify the way in which users can gain reputation, and thus define, and prescribe, what constitutes "good behavior" on the users' part. Users are strongly encouraged to follow the prescribed behavior, lest their reputation — and their ability to use the system — suffers.

- *Descriptive value.* Reputation systems can be used to classify users, and their contributions, on the basis of their reputation, making it easier to spot high-quality users or contributions, and flagging contributions or users with low reputation.

Ebay's reputation system is an example of a system that has both prescriptive, and descriptive, value. Users are strongly encouraged to accumulate positive feedback: it has been documented that positive feedback increases the probability of closing a transaction, and in the case of sellers, is connected to higher selling price for goods [LRBPR99]. The original PageRank algorithm [PBMW98] constitutes a reputation system for web pages with descriptive intent.

Our reputation system for the Wikipedia has prescriptive and descriptive value. Reputation is built by contributing lasting content; by flagging author reputation, we encourage such contributions. The reputation we compute also has descriptive value: Wikipedia visitors can use the author's reputation as a guide to the trustworthiness of freshly-contributed text.

In addition to prescriptive and descriptive values, we argue that a reputation system that is truly useful for the Wikipedia must also have *predictive* value: an author's reputation should be statistically related to the quality of the author's *future* contributions. To a Wikipedia visitor, it is not generally very useful to know the reputation of an author who made a long-past contribution to an article. If the contribution survived for a long time, its quality is essentially proven: all subsequent editors to the same page have already implicitly voted on the contribution by leaving it in place. Reputation in the Wikipedia is most useful as a guide to the value of fresh contributions, which have not yet been vetted by other authors. Reputation is most useful if it can predict how well these fresh contributions will fare; whether they are likely to be long-lasting, or whether they are likely to be reverted in short order. Furthermore, when reputation is used to grant or deny the ability to edit a page, it is its predictive value that matters: after all, why deny the ability to edit a page, unless there is some statistical reason to believe that the edits will need to be undone?

1.3 Summary of the Results

In order to measure the predictive value of the proposed reputation, we implemented our reputation system, and we used it to analyze all edits done to the Italian Wikipedia from its inception to October 31, 2005 (154,621 pages, 714,280 versions), and all edits done to the French Wikipedia from its inception, to July 30, 2006 (536,930 pages, 4,837,243 versions).[3] We then measured the statistical correlation between the author's reputation at the time an edit was done, and the subsequent lifespan of the edit. This is a valid statistical test, in the sense that the two variables of the study are computed in independent ways: the reputation of an author at the time of the edit is computed from the behavior of the author *before* the edit, while the lifespan of the edit is determined by events *after* the edit. To summarize the results, we introduce the following informal terminology; we will make the terms fully precise in the following sections:

- *Short-lived edit* is an edit whose effect is undone in the next couple of revisions;

- *Short-lived text* is text that is almost immediately removed (each successive revision removes, on average, 4/5 of such text).

- *Low-reputation author* is an author whose reputation falls in the bottom 20% of the reputation scale.

When measuring the quantity of text, or edits, our unit of measurement is a word (white-space delimited string): this provides a uniform unit of measurement, and ensures that splitting changes in two or more revisions does not affect the results. This also explains why the percentages of text and edits done by low-reputation authors are slightly different. Our results for the French Wikipedia indicate that the fact that an author's reputation is low at the time a contribution is made is statistically correlated with the contribution being undone:

- 7.7% of the edits are performed by low-reputation authors; these 7.7% edits account for 32% of the short-lived edits. Edits by low-reputation authors are 4.2 times more likely than average to be short-lived.

- 8.4% of the contributed text comes from low-reputation authors; this 8.4% text accounts for 38%

[3] These numbers reflect the fact that we consider only the last among consecutive versions by the same author, as explained later.

of the short-lived text. Text contributed by low-reputation users is 4.5 times more likely than average to be short-lived.

Using search terminology, the *recall* provided by low reputation is 32% for short-lived edits and 38% for short-lived text. The *precision* is as follows:

- 24% of the edits done by low-reputation authors are short-lived.
- 5.8% of the text inserted by low-reputation authors is short-lived.

The difference in these two precision figures indicates that most edits performed by low-reputation authors that are short-lived do not correspond to much insertion of new text (they may involve text removal or displacement, instead). While we use search terminology to report these results, we stress that these results are about the reputation's ability to *predict* the longevity of future contributions by an author.

The above results may underestimate the recall of our content-driven reputation. We asked a group of volunteers to decide, of the short-lived contributions to the Italian Wikipedia, which ones were of poor quality. The results indicated that short-lived contributions by low-reputation authors were markedly more likely to be of poor quality, compared to similarly short-lived contributions by high-reputation authors. This allowed us to calculate that the recall for bad-quality, short-lived edits is 49%, and the recall for bad-quality short-lived text is 79%.

We are unsure of the extent with which prediction precision can be increased. Many authors with low reputation are good, but novice, contributors, who have not had time yet to accumulate reputation. Indeed, an unfortunate effect of the ease with which users can register anew to the Wikipedia is that we cannot trust novices any more than confirmed bad contributors — if we trusted novices more, bad contributors would simply re-register. Furthermore, even authors who contribute short-lived text and edits, and who therefore have low reputation, do not do so consistently, interjecting good contributions among the bad ones.

A basic form of reputation, currently used to screen contributors to controversial pages, is the length of time for which users have been registered to the Wikipedia. We did not have access to this temporal information in our study.[4] As a proxy, we considered the number of edits performed by a user. This is a form of reputation that has no prescriptive value: all it does is encouraging users to split their contributions in multiple, small ones, or even worse, to perform a multitude of gratuitous edits. Indeed, if edit count were used as reputation, it would most likely induce undesirable behaviors by the Wikipedia users. We will show that the predictive value of such a reputation is also inferior to the one we compute, although the difference is not large, and we will discuss why we believe this is the case.

1.4 Related Work

The work most closely related to ours is [ZAD$^+$06], where the revision history of a Wikipedia article is used to compute a trust value for the article; dynamic Bayesian networks are used to model the evolution of trust level over the versions. At each revision, the inputs to the network are a priori models of trust of authors (determined by their Wikipedia ranks), and the amount of added and deleted text. The paper shows that this approach can be used to predict the quality of an article; for instance, it can be used to predict when an article in a test set can be used as a featured article. Differently from the current work, author trustworthiness is taken as input; we compute author reputation as output. Furthermore, the work tracks the amount of added and deleted text at each revision, but does not track whose text it is, nor does it track text across revisions. Thus, it cannot be used to infer the trustworthiness of authors. A simpler approach to text trust, based solely on text age, is described in [Cro06]. In many ways, estimating text trust, and author reputation, are complementary approaches. Text trust works best for text that has been part of an article for some time; author reputation is most useful as an indicator of quality for fresh text.

Reputation systems in e-commerce and social networks have been extensively studied (see, e.g., [Kle99, RZFK00, Del03, KSGM03]); the reputation in those systems is generally user-driven, rather than content-driven as in our case. Related is also work on trust in social networks (see, e.g., [GKRT04, Gol05]).

The history flow of text contributed by Wikipedia authors has been studied with flow visualization methods in [VWD04]; the results have been used to analyze a number of interesting patterns in the content evolution of Wikipedia articles. Work on mining software revision logs (see, e.g., [LZ05]) is similar in its emphasis of in-depth analysis of revision logs; the aim, however, is to find revision patterns and indicators that point to software defects, rather than to develop a notion of author reputation.

[4] Wikipedia makes only a limited amount of user information available for download.

2 Content-Driven Reputation

Reputation systems can be classified into two broad categories: *chronological,* where the reputation is computed from the chronological sequence of ratings a user receives, and *fixpoint,* where the reputation is computed via a fixpoint calculation performed over the graph of feedbacks. The Ebay reputation system is an example of a chronological system, while the PageRank and HITS algorithms are examples of fixpoint algorithms [PBMW98, Kle99]. We chose to follow the chronological approach to develop our content-driven reputation for the Wikipedia. The chief advantage of a chronological approach is that it is computationally lightweight. When an author revises a Wikipedia article, we can efficiently compute the feedback to authors of previous revisions to the same article, and we can modify their reputation in real-time, with little impact on server response time.

2.1 Notation

The following notation will be used throughout the paper. We assume that we have $n > 0$ versions $v_0, v_1, v_2, \ldots, v_n$ of a document; version v_0 is empty, and version v_i, for $1 \leq i \leq n$, is obtained by author a_i performing a revision $r_i : v_{i-1} \rightsquigarrow v_i$. We refer to the change set corresponding to $r_i : v_{i-1} \rightsquigarrow v_i$ as the *edit* performed at r_i: the edit consists of the text insertions, deletions, displacements, and replacements that led from v_{i-1} to v_i. When editing a versioned documents, authors commonly save several versions in a short time frame, in order to avoid losing their work in case of computer or network problems. To ensure that such behavior does not affect reputations, we *filter* the versions, keeping only the *last* of consecutive versions by the same author; we assume thus that for $1 \leq i < n$ we have $a_i \neq a_{i+1}$. Every version v_i, for $0 \leq i \leq n$, consists of a sequence $[w^i_1, \ldots, w^i_{m_i}]$ of words, where m_i is the number of words of v_i; we have $m_0 = 0$. For us, a *word* is a whitespace-delimited sequence of characters in the Wiki markup language that produces a Wikipedia article: we work at the level of such markup language, rather than at the level of the HTML produced by the wiki engine. Given a series of versions $v_0, v_1, \ldots, v_n$ of a text document, we assume that we can compute the following quantities:

- $txt(i, j)$, for $0 < i \leq j \leq n$, is the amount of text (measured in number of words) that is introduced by r_i in v_i, and that is still present (and due to the same author a_i) in v_j. In particular, $txt(i, i)$ is the amount of new text added by r_i.
- $d(v_i, v_j)$, for $0 < i < j \leq n$, is the *edit distance* between v_i and v_j, and measures how much change (word additions, deletions, replacements, displacements, etc.) there has been in going from v_i to v_j.

We will describe in the next section how these quantities can be computed from the series of versions $v_0, \ldots, v_n$.

2.2 Content-Driven Reputation in a Versioned Document

We propose the following method for computing content-driven reputation in a versioned document. Consider a revision $r_i : v_{i-1} \rightsquigarrow v_i$, performed by user a_i, for some $0 < i \leq n$. Each of the subsequent authors $a_{i+1}, a_{i+2}, \ldots$ can either retain, or remove, the edits performed by a_i at r_i; in particular, they can either keep, or discard, the text introduced by r_i. We then examine later versions v_j of the document, for $i < j \leq n$, and we increase (or decrease) the reputation of a_i according to how much of the new text, and edits, introduced in r_i have survived until v_j. In the evaluation, we also take into account the reputation of author a_j of v_j, who acts as the judge. This will ensure that high-reputation authors do not risk much of their reputation while fighting revision wars against very low reputation authors, such as anonymous authors. If this were not the case, high-reputation authors would be wary of undoing damage done by such low-reputation authors, including spammers, for fear of reprisal.

We use two criteria to award reputation: the survival of text, and the survival of edits. Text survival is perhaps the most obvious criterion. Adding new text to a Wikipedia article is a fundamental way of contributing — it is how new knowledge is added — and it seems reasonable to take into account the amount of text added, and its longevity, when computing author reputation. However, text survival alone fails to capture many ways in which authors contribute to the Wikipedia. If a user rearranges the content of an article without introducing new text, the contribution cannot be captured by text survival. The act of restoring an article to a previous, good version after an act of vandalism does not result in the addition of new text; thus, if we based reputation on text survival alone, authors who undo damage and restore articles would not gain any reputation from their actions. This, in spite of the fact that such repairs are fundamental to the quality of an open wiki. Edit survival captures how long the re-arrangements performed by an author last in the history of a Wikipedia article, and captures all of the above ways of contributing.

We use a parameter $c_{text} \in [0, 1]$ to determine how much weight should be given to text versus edit survival. We have determined the parameter c_{text}, along with other parameters used in the calculation, via an optimization process aimed at producing reputation functions of high predictive value; the optimization process is described later in Section 4.

2.3 Accounting for Text Survival

If text introduced by r_i is still present in v_j, for $0 < i < j \leq n$, this indicates that author a_j, who performed the revision $r_j : v_{j-1} \rightsquigarrow v_j$, agrees that the text is valuable. To reflect this, we increase the reputation of a_i, in a manner that is proportional to the amount of residual text, and to the reputation of author a_i. Precisely, the rule we have chosen is as follows.

Rule 1 (update due to text survival) *When the revision r_j occurs, for all $0 < i < j$ such that $j - i \leq 10$ and $a_j \neq a_i$, we add to the reputation of a_i the amount:*

$$c_{scale} \cdot c_{text} \cdot \frac{txt(i,j)}{txt(i,i)} \cdot (txt(i,i))^{c_{len}} \cdot \log(1 + R(a_j, r_j)),$$

where $c_{scale} > 0$, $c_{text} \in [0,1]$, and $c_{len} \in [0,1]$ are parameters, and where $R(a_j, r_j)$ is the reputation of a_j at the time r_j is performed.

The rule can be understood as follows. $txt(i,j)/txt(i,i)$ is the fraction of text introduced at version v_i that is still present in version v_j; this is a measure of the "quality" of r_i. The quantity $\log(1 + R(a_j, r_j))$ is the "weight" of the reputation of a_j, that is, how much the reputation of a_j lends credibility to the judgements of a_j. We chose to adopt a logarithmic notion of weight, because in our reputation systems, the reputations of most regular contributors, who rarely perform low-quality edits, quickly soar to very high values compared to the reputations of beginners (this will be evident from the experimental results we present later). Using a logarithmic weight for reputations limits the power of established authors, preventing them from entirely overriding feedback coming from newer members. The parameters c_{scale}, c_{text} and c_{len} will be determined experimentally via an optimization process, as described later. Their meaning is as follows:

- The parameter $c_{len} \in [0,1]$ is an exponent that specifies how to take into account the length of the original contribution: if $c_{len} = 1$, then the increment is proportional to the length of the original contribution; if $c_{len} = 0$, then the increment does not depend on the length of the original contribution.
- The parameter c_{scale} specifies how much the reputation should vary in response to an individual feedback.

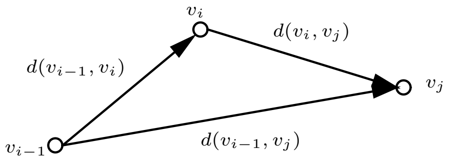

Figure 1: Distances involved in the computation of $ELong(i,j)$.

- The parameter c_{text} specifies how much the feedback should depends on residual text (Rule 1) or residual edit (Rule 2, presented later).

In the rule, only the 10 most recent versions of an article are considered. This ensures that contributors to articles that have a natural, slow rate of change get roughly the same feedback of contributors to articles that reach a steady, or even frozen, state. We considered basing the limit on time, rather than number of versions, but each Wikipedia article has its own rate of change, dictated in part by the popularity of the article, so we believe that the number of versions is a more uniform criterion.

2.4 Accounting for Edit Survival

To judge the revision $r_i : v_{i-1} \rightsquigarrow v_i$ from the vantage point of v_j, for $0 < i < j \leq n$, we reason as follows. We wish to rate r_i higher, if r_i made the article more similar to v_j. In particular, the revision r_i changed the article by an amount $d(v_{i-1}, v_i)$; we wish to credit a_i in proportion to how much of this change is directed towards v_j. This suggests using the formula:

$$ELong(i,j) = \frac{d(v_{i-1}, v_j) - d(v_i, v_j)}{d(v_{i-1}, v_i)}. \qquad (1)$$

Figure 1 depicts the distances involved in the computation of $ELong(i,j)$. If d satisfies the triangular inequality (as our edit distance does, to a high degree of accuracy), then $ELong(i,j) \in [-1,1]$. For two consecutive edits r_i, r_{i+1}, if r_i is completely undone in r_{i+1} (as is common when r_i consists in the introduction of spam or inappropriate material), then $ELong(i,i+1) = -1$; if r_{i+1} entirely preserves r_i, then $ELong(i,i+1) = +1$. For more distant edits, $ELong(i,j)$ is a measure of how much of the edit performed during r_i is undone (value -1) or preserved (value $+1$) before r_j. The rule we use for updating the reputation is as follows.

Rule 2 (update due to edit survival) *When the revision r_j occurs, for all $0 < i < j$ such that $j - i \leq 3$, we add to the reputation of a_i an amount q determined as follows. If $a_j = a_i$ or $d(v_{i-1}, v_i) = 0$, then $q = 0$; otherwise, q is determined by the following algorithm.*

$$q := \frac{c_{slack} \cdot d(v_{i-1}, v_j) - d(v_i, v_j)}{d(v_{i-1}, v_i)}$$

if $q < 0$ **then** $q := q \cdot c_{punish}$ **endif**

$$q := q \cdot \Big(c_{scale} \cdot (1 - c_{text}) \cdot \big(d(v_{i-1}, v_i)\big)^{c_{len}} \cdot \log\big(1 + R(a_j, r_j)\big)\Big)$$

In the algorithm, $c_{punish} \geq 1$*,* $c_{slack} \geq 1$*,* $c_{scale} > 0$*,* $c_{text} \in [0, 1]$*, and* $c_{len} \in [0, 1]$ *are parameters, and* $R(a_j, r_j)$ *is the reputation of* a_j *at the time* r_j *is performed.*

The rule adopts a modified version of (1): the parameter $c_{slack} > 1$ is used to allow revision $r_i : v_{i-1} \rightsquigarrow v_i$ to be slightly counterproductive, without causing punishment for a_i. On the other hand, when punishment is incurred, the parameter c_{punish} is used to amplify its magnitude, compared to the amount of positive reputation gained from good edits. Amplifying punishment is instrumental to make the threat a credible one. Without amplification, a rogue contributor could use the reputation gained in one part of the Wikipedia to constantly destroy a small set of articles elsewhere. Amplification makes this harder to achieve. The parameters c_{slack} and c_{punish}, as well as c_{scale}, c_{text} and c_{len}, will be determined via an optimization process, as mentioned before. To assign edit feedback, we have chosen consider only the 3 previous versions of an article. This approach proved adequate for analyzing an already-existing wiki, in which authors could not modify their behavior using knowledge of this threshold. If the proposed content-driven reputation were to be used on a live Wiki, it would be advisable to replace this hard threshold by a scheme in which the feedback of v_j on r_i is weighed by a gradually decreasing function of $j - i$ (such as $\exp(c \cdot (i - j))$ for some $c > 0$).

2.5 Computation of Content-Driven Reputation

We compute the reputation for Wikipedia authors as follows. We examine all revisions in chronological order — thus simulating the same order in which they were submitted to the Wikipedia servers. We initialize the reputations of all authors to the value 0.1; the reputation of anonymous authors is fixed to 0.1. As the revisions are processed, we use Rules 1 and 2 to update the reputations of authors in the system. When updating reputations, we ensure that they never become negative, and that they never grow beyond a bound $c_{maxrep} > 0$. The bound c_{maxrep} is used to prevent frequent contributors from accumulating unbounded amounts of reputation, and becoming essentially immune to negative feedback. The value of c_{maxrep} will be once again determined via optimization techniques.

Wikipedia allows users to register, and create a *author* identity, whenever they wish. As a consequence, we need to make the initial reputation of new authors very low, close to the minimum possible (in our case, 0). If we made the initial reputation of new authors any higher, then rogue authors, after committing revisions that damage their reputation, would simply re-register as new users to gain the higher value. An unfortunate side-effect of allowing people to obtain new identities at will is that we cannot presume that people are innocent until proven otherwise: we have to assign to newcomers the same reputation as proven offenders. This is a contributing factor to our reputation having low *precision:* many authors who have low reputation still perform very good quality revisions, as they are simply new authors, rather than proven offenders. We conjecture that content-driven reputation systems for the Wikipedia would have better predictive value if creating a new author identity was not free, either monetarily, or in some other sense.

3 Text Longevity and Edit Distance in Versioned Documents

In this section, we describe in more detail how we tracked text authorship, and how we computed edit distances, in versioned documents. We developed our algorithms starting from standard text-difference algorithms, and in particular, those of [Tic84, Mye86, BL97]. However, we adapted the algorithms in several places, to enable them to better cope with the versioned nature of the Wikipedia, and with the kind of edits that authors perform.

3.1 Tracking Text Authorship

Given a sequence of versions $v_0, v_1, \ldots, v_n$, we describe an algorithm for computing $txt(i, j)$ for $0 < i \leq j \leq n$. Superficially, it might seem that to compute $txt(i, j)$, we need to consider only three versions of the document: v_{i-1}, v_i, and v_j. From v_{i-1} and v_i we can derive the text that is added at revision $r_i : v_{i-1} \rightsquigarrow v_i$, and we can then check how much of it survives until v_j. This approach, however, is not appropriate for a versioned document like a wiki page, where authors are allowed to inspect— and restore — text from any previous version of a document. For example, consider the case in which revision $r_{i-1} : v_{i-2} \rightsquigarrow v_{i-1}$ is the work of a spammer, who erases entirely the text of v_{i-2}, and replaces it with spurious material; such spam insertions are a common occurrence in open wikis. When author a_i views the page, she realizes that it has been damaged, and she reverts it to the previous version, so that $v_i = v_{i-2}$. If we derived the text added by r_i by considering v_{i-1} and v_i only, it would appear to us that a_i is the original author of all the text in v_i, but this is clearly not the case: she simply restored pre-existing text. To compute the text added at a revision $r_i : v_{i-1} \rightsquigarrow v_i$, we keep track of both the text that is in v_{i-1}, and of the text that used to be present in previous versions, and that has subsequently been deleted.

Our algorithm proceeds as follows. We call a *chunk* a list $c = [(w_1, q_1), \ldots, (w_k, q_k)]$, where for $1 \leq j \leq k$, w_j is a word, and $q_j \in \{1, \ldots, n\}$ is a version number. A chunk represents a list of contiguous words, each labeled with the version where it originates. The algorithm computes, for each version v_i, $1 \leq i \leq n$, its *chunk list* $C_i = [c_0^i, c_1^i, \ldots, c_k^i]$. The chunk c_0^i is the *live* chunk, and it consists of the text of v_i, with each word labeled with the version where the word was introduced; thus, if $v_i = [w_1^i, \ldots, w_{m_i}^i]$, we have $c_0^i = [(w_1, q_1), \ldots, (w_{m_i}, q_{m_i})]$, for some $q_1, \ldots, q_{m_i}$. The chunks $c_1^i, \ldots, c_k^i$ are *dead* chunks, and they represent contiguous portions of text that used to be present in some version of the document prior to i. Given the chunk list $C_i = [c_0^i, c_1^i, \ldots, c_k^i]$ for document v_i, we can compute $txt(j, i)$ via

$$txt(j,i) = \left|\left\{(u,j) \;\middle|\; \exists u.(u,j) \in c_0^i\right\}\right|, \tag{2}$$

for $1 \leq i \leq j \leq n$.

To compute C_i for all versions i of a document, we propose an algorithm that proceeds as follows. For the initial version, we let $C_1 = [[(w_1^1, 1), (w_2^1, 1), \ldots, (w_{m_1}^1, 1)]]$. For $1 \leq i < n$, the algorithm computes C_{i+1} from $C_i = [c_0^i, c_1^i, \ldots, c_k^i]$ and v_{i+1}. To compute the live chunk c_0^{i+1}, we match contiguous portions of text in v_{i+1} with contiguous text in any of the chunks in C_i; the matching words in v_{i+1} are labeled with the version index that labels them in C_i, and represent words that were introduced prior to version v_{i+1}. Any words of v_{i+1} that cannot be thus matched are considered new in v_{i+1}, and are labeled with version index $i+1$. The dead chunks $c_1^{i+1}, \ldots, c_l^{i+1}$ of C_{i+1} are then obtained as the portions of the chunks in C_i that were not matched by any text in v_{i+1}. We allow the same text in C_i to be matched multiple timed in v_{i+1}: if a contributor copies multiple times text present in v_i or in prior versions in order to obtain v_{i+1}, the replicated text should not be counted as new in v_{i+1}. Considering replicated text as new would open the door to a *duplication attack,* whereby an attacker duplicates text in a revision r_i, and then removes the original text in a revision $r_k : v_{k-1} \rightsquigarrow v_k$ with $k > i$. From version v_k onwards, the text would be attributed to the attacker rather than to the original author.

Matching v_{i+1} with C_i is a matter of finding matches between text strings, and several algorithms have been presented in the literature to accomplish this in an efficient manner (see, e.g., [HM75, Hir77, Tic84, Mye86, BL97]). We experimented extensively, and the algorithm that gave the best combination of efficiency and accuracy was a variation of a standard greedy algorithm. In standard greedy algorithms, such as [Hir77, Mye86, BL97], longest matches are determined first; in our algorithm, we define a notion of match *quality,* and we determine first matches of highest quality. To define match quality, we let m_{i+1} be the length of v_{i+1}, and we let m' be the length of the chunk of C_i where the match is found (all length and indices are measured in number of words). Let l be the length of the match, and assume that the match begins at word $k' \leq m'$ in the chunk, and at word $k_{i+1} \leq m_{i+1}$ in v_{i+1}. We define match quality as follows:

- If the match occurs between v_{i+1} and the live chunk, then the quality is:

$$\frac{l}{\min(m_{i+1}, m')} - 0.3 \cdot \left| \frac{k'}{m'} - \frac{k_{i+1}}{m_{i+1}} \right| .$$

- If the match occurs between v_{i+1} and a dead chunk, then the quality is 0 if $l < 4$, and is $l/\min(m_{i+1}, m') - 0.4$ otherwise.

Thus, the quality of a match is the higher, the longer the match is. If the match is with the live chunk, a match has higher quality if the text appears in the same relative position in v_{i+1} and in v_i. Matches with dead chunks have somewhat lower quality than matches with the live chunk: this corresponds to the fact that, if some text can be traced both to the previous version (the live chunk), and to some text that was previously deleted, the most likely match is with the text of the previous version. Moreover, matches with dead chunks have to be at least of length 4: this avoids misclassifying common words in new text as re-introductions of previously-deleted text. The coefficients in the above definition of quality have been determined experimentally, comparing human judgements of authorship to the algorithmically computed ones for many pages of the Italian Wikipedia. The text survival algorithm we developed is efficient: the main bottleneck, when computing text authorship, is not the running time of the algorithm, but rather, the time required to retrieve all versions of a page from the MySQL database in which Wikipedia pages are stored.[5]

[5] The measurement was done on a PC with AMD Athlon 64 3000+ CPU, two hard drives configured in RAID 1 (mirroring), and 1 GB of memory.

3.2 Computing Edit Distances

For two versions v, v', the edit distance $d(v, v')$ is a measure of how many word insertions, deletions, replacements, and displacements are required to change v into v'. Wikipedia pages contain markup, and possibly the most accurate way to measure edit distance would treat page versions as structured documents, and apply algorithms for the edit distance of structured documents [ZS89, CGM97, CAM02]. We opted however for using a standard notion of edit distance for flat files [Tic84]. In order to compute the edit distance between two versions v and v', we use the same greedy algorithm for text matching that we used for text survival, except that each portion of text in v (resp. v') can be matched *at most once* with a portion of text in v' (resp. v). Thus, text duplication is captured as an edit. The output of the greedy matching is modified, as is standard in measurements of edit distance, so that it outputs a list L of elements that describe how v' can be obtained from v:

- $I(j, k)$: k words are inserted at position j;
- $D(j, k)$: k words are deleted at position j;
- $M(j, h, k)$: k words are moved from position j in v to position h in v'.

We compute the total amount I_{tot} of inserted text by summing, for each $I(j, k) \in L$, the length k; similarly, we obtain the total amount D_{tot} of deleted text by summing, for each $D(j, k) \in L$, the length k. We take into account insertions and deletions via the formula $I_{tot} + D_{tot} - \frac{1}{2}\min(I_{tot}, D_{tot})$: thus, every word that is inserted or removed contributes 1 to the distance, and every word that is replaced contributes $\frac{1}{2}$. The motivation for this treatment of replacements is as follows. Suppose that author a_i edits v_{i-1} adding a new paragraph consisting of k words, obtaining v_i, and suppose that author a_{i+1} then rewrites completely the paragraph (keeping it of equal length), obtaining v_{i+1}. If we accounted for insertion and deletions via $I_{tot} + D_{tot}$, then $d(v_{i+1}, v_i) = 2k$, while $d(v_{i+1}, v_{i-1}) = k$: the edit performed by author a_i would thus be considered highly counterproductive. With our formula, we have instead $d(v_{i+1}, v_i) = k/2$ and $d(v_{i+1}, v_{i-1}) = k$, so that the contribution of author a_i is positive: experiments we performed showed that, on average, this is in better agreement with a human perception of the contribution of author a_i.

We account for text moves between versions v and v' as follows. Let l, l' be the lengths (in words) of v and v', respectively. Each time a block of text of length k_1 exchanges position with a block of text of length k_2, we count this as $k_1 \cdot k_2 / \max(l, l')$. Thus, a word that moves across k other words contributes $k/\max(l, l')$ to the distance: the contribution approaches 1 as the word is moved across the whole document. The total contribution M_{tot} of all moves can be computed by adding $k \cdot k'$, for all pairs of moves $M(j, h, k) \in L$ and $M(j', h', k') \in L$ such that $j < j'$ and $h > h'$ (this ensures that every crossing is counted once). We finally define:

$$d(r, r') = I_{tot} + D_{tot} + M_{tot} - \tfrac{1}{2}\min(I_{tot}, D_{tot}).$$

Due to the nature of the greedy algorithms used for text matching, and of the definitions above, our edit distance is not guaranteed to satisfy the triangular inequality. However, we found experimentally that the proposed edit distance, on Wikipedia pages, satisfies the triangular inequality within approximately one unit (one word) for well over 99% of triples of versions of the same page.

4 Evaluation and Optimization Metrics

We now develop quantitative measures of the ability of our content-driven reputation to predict the quality of future revisions. For a revision $r_i : v_{i-1} \rightsquigarrow v_i$ in a sequence $v_0, v_1, \ldots, v_n$ of versions, let $\rho_t(r_i) = \mathit{txt}(i, i)$ be the new text introduced at r_i, and $\rho_e(r_i) = d(v_{i-1}, v_i)$ be the amount of editing involved in r_i. We define edit and text longevity as follows:

- The *edit longevity* $\alpha_e(r_i) \in [-1, 1]$ of r_i is the average of $\mathit{ELong}(i, j)$ for $i < j \le \min(i + 3, n)$.
- The *text longevity* $\alpha_t(r_i) \in [0, 1]$ of r_i is the solution to the following equation:

$$\sum_{j=i}^{n} \mathit{txt}(i, j) = \mathit{txt}(i, i) \cdot \sum_{j=1}^{n} (\alpha_t(r_i))^{j-i} . \qquad (3)$$

Thus, $\alpha_t(r_i)$ is the coefficient of exponential decay of the text introduced by r_i: on average, after k revisions, only a fraction $(\alpha_t(r_i))^k$ of the introduced text survives. As all quantities in (3) except $\alpha_t(r_i)$ are known, we can solve for $\alpha_t(r_i)$ using standard numerical methods. We also indicate by $\mathit{rep}(r_i)$ the reputation of the author a_i of r_i at the time r_i was performed. We note that $\mathit{rep}(r_i)$ is computed from the history of the Wikipedia before r_i, while $\alpha_e(r_i)$ and $\alpha_t(r_i)$ depend only on events after r_i. Moreover, $\alpha_e(r_i)$ and $\alpha_t(r_i)$ can be computed independently of reputations.

Let R be the set of all revisions in the Wikipedia (of all articles). We view revisions as a probabilistic process, with R as the set of outcomes. Since some revisions change very little, while others affect many words, we normalize revisions by the number of words they affect. This ensures that the metrics are not affected if revisions by the same author are combined or split in multiple steps. Since we keep only the last among consecutive revisions by the same user, a "revision" is a rather arbitrary unit of measurement, while a "revision amount" provides a better metric.

Thus, when studying edit longevity, we associate with each $r \in R$ a probability mass proportional to $\rho_e(r)$, giving rise to the probability measure $\Pr_e$. Similarly, when studying text longevity, we associate with each $r \in R$ a probability mass proportional to $\rho_t(r)$, giving rise to the probability measure $\Pr_t$.

In order to develop figures of merit for our reputation, we define the following terminology (used already in the introduction in informal fashion):

- We say that the edit performed in r is *short-lived* if $\alpha_e(r) \leq -0.8$.
- We say that the new text added in r is *short-lived* if $\alpha_t(r) \leq 0.2$, indicating that at most 20% of it, on average, survives from one version to the next.
- We say that a revision r is *low-reputation* if $\log(1 + \mathit{rep}(r)) \leq \log(1 + c_{maxrep})/5$, indicating that the reputation, after logarithmic scaling, falls in the lowest 20% of the range.

Correspondingly, we define three random variables $S_e, S_t, L : R \mapsto \{0, 1\}$ as follows, for all $r \in R$:

- $S_e(r) = 1$ if $\alpha_e(r) \leq -0.8$, and $S_e(r) = 0$ otherwise.
- $S_t(r) = 1$ if $\alpha_e(r) \leq 0.2$, and $S_t(r) = 0$ otherwise.
- $L(r) = 1$ if $\log(1 + \mathit{rep}(r)) \leq \log(1 + c_{maxrep})/5$, and $L(r) = 0$ otherwise.

The *precision* prec_t and *recall* rec_t for short-lived text, and the *precision* prec_e and *recall* rec_e for short-lived edits, are defined as:

$$\mathit{prec}_t = \Pr_t(S_t{=}1 \mid L{=}1) \qquad \mathit{rec}_t = \Pr_t(L{=}1 \mid S_t{=}1)$$
$$\mathit{prec}_e = \Pr_e(S_e{=}1 \mid L{=}1) \qquad \mathit{rec}_e = \Pr_e(L{=}1 \mid S_e{=}1).$$

These quantities can be computed as usual; for instance,

$$\Pr_e(S_e = 1 \mid L = 1) = \frac{\sum_{r \in R} S_e(r) \cdot L(r) \cdot \rho_e(r)}{\sum_{r \in R} L(r) \cdot \rho_e(r)}$$

(noting that it is unnecessary in these expressions to renormalize all probability masses via $1/\sum_{r \in R} \rho_e(r)$). We also define:

$$\mathit{boost}_e = \frac{\Pr_e(S_e{=}1 \mid L{=}1)}{\Pr_e(S_e{=}1)} = \frac{\Pr_e(S_e{=}1, L{=}1)}{\Pr_e(S_e{=}1) \cdot \Pr_e(L{=}1)}$$
$$\mathit{boost}_t = \frac{\Pr_t(S_t{=}1 \mid L{=}1)}{\Pr_t(S_t{=}1)} = \frac{\Pr_t(S_t{=}1, L{=}1)}{\Pr_t(S_t{=}1) \cdot \Pr_t(L{=}1)}$$

Reputation	Judged bad	Judged good
Short-lived edits:		
Low [0.0–0.2]	66 %	19 %
Normal [0.2–1.0]	16 %	68 %
Short-lived text:		
Low [0.0–0.2]	74 %	13 %
Normal [0.2–1.0]	14 %	85 %

Table 2: User ranking of short-lived edits and text, as a function of author reputation. In square brackets, we give the interval where the normalized value $\log(1 + r)/\log(1 + c_{maxrep})$ of a reputation r falls. The precentages do not add to 100%, because users could also rank a change as "neutral".

Intuitively, boost_e indicates how much more likely than average it is that edits produced by low-reputation authors are short-lived. The quantity boost_t has a similar meaning. Our last indicator of quality are the *coefficients of constraint*

$$\kappa_e = I_e(S_e, L)/H_e(L) \qquad \kappa_t = I_t(S_t, L)/H_t(L),$$

where I_e is the *mutual information* of S_e and L, computed with respect to $\Pr_e$, and H_e is the entropy of L, computed with respect to $\Pr_e$ [CT91]; similarly for $I_t(S_t, L)$ and $H_t(L)$. The quantity κ_e is the fraction of the entropy of the edit longevity which can be explained by the reputation of the author; this is an information-theoretic measure of correlation. The quantity κ_t has an analogous meaning.

To assign a value to the coefficients c_{scale}, c_{slack}, c_{punish}, c_{text}, c_{len}, and c_{maxrep}, we implemented a search procedure, whose goal was to find values for the parameters that maximized a given objective function. We applied the search procedure to the Italian Wikipedia, reserving the French Wikipedia for validation, once the coefficients were determined. We experimented with both κ_e and $\mathit{prec}_e \cdot \mathit{rec}_e$ as objective functions, and they both gave very similar results.

5 Experimental Results

To evaluate our content-driven reputation, we considered two Wikipedias:

- The Italian Wikipedia, consisting of 154,621 articles and 714,280 *filtered* revisions; we used a snapshot dated December 11, 2005.
- The French Wikipedia, consisting of 536,930 articles and 4,837,243 *filtered* revisions; we used a snapshot dated October 14, 2006.

	Precision		Recall		Boost		Coeff. of constr.	
	Edit	Text	Edit	Text	Edit	Text	Edit	Text
	$prec_e$	$prec_t$	rec_e	rec_t	$boost_e$	$boost_t$	κ_e	κ_t
Italian Wikipedia	14.15	3.94	19.39	38.69	4.03	5.83	3.35	7.17
French Wikipedia	23.92	5.85	32.24	37.80	4.21	4.51	7.33	6.29

Table 1: Summary of the performance of content-driven reputation over the Italian and French Wikipedias. All data are expressed as percentages.

Of both wikipedias, we studied only "NS_MAIN" pages, which correspond to ordinary articles (other pages are used as comment pages, or have other specialized purposes). Moreover, to allow for the more accurate computation of edit longevity, we used only revisions that occurred before October 31, 2005 for the Italian Wikipedia, and before July 31, 2006 for the French one. Our algorithms for computing content-driven reputation depend on the value of six parameters, as mentioned earlier. We determined values for these parameters by searching the parameter space to optimize the coefficient of constraint, using the Italian Wikipedia as a training set; the values we determined are:

$$c_{scale} = 13.08 \quad c_{punish} = 19.09 \quad c_{len} = 0.60$$
$$c_{slack} = 2.20 \quad c_{text} = 0.60 \quad c_{maxrep} = 22026$$

We then analyzed the Italian and French Wikipedias using the above values. The results are summarized in Table 1. The results are better for the larger French Wikipedia; in particular, the reputation's ability to predict short-lived edits is better on the French than on the Italian Wikipedias. We are not sure whether this depends on different dynamics in the two Wikipedias, or whether it is due to the greater age (and size) of the French Wikipedia; we plan to study this in further work. We see that edits performed by low-reputation authors are four times as likely as the average to be short-lived.

To investigate how many of the edits had a short life due to bad quality, we asked a group of 7 volunteers to rate revisions performed to the Italian Wikipedia. We selected the revisions to be ranked so that they contained representatives of all 4 combinations of high/low reputation author, and high/low longevity. We asked the volunteers to rate the revisions with +1 (good), 0 (neutral), and −1 (bad); in total, 680 revisions were ranked. The results, summarized in Table 2, are striking. Of the short-lived edits performed by low-reputation users, fully 75% were judged bad. On the other hand, less than 9.2% of the short-lived edits performed by high-reputation users were judged bad. We analyzed in detail the relationship between user reputation, and the percentage of short-lived text and edits that users considered bad. Using these results, we computed the approximate recall factors on the Italian Wikipedia of content-driven reputation for *bad* edits, as judged by users, rather than short-lived ones:

- The recall for short-lived edits that are judged to be bad is over 49%.
- The recall for short-lived text that is judged to be bad is over 79%.

These results clearly indicate that our content-driven reputation is a very effective tool for spotting, at the moment they are introduced, bad contributions that will later be undone. There is some margin of error in this data, as our basis for evaluation is a small number of manually-rated revisions, and human judgement on the same revisions often contained discrepancies. We note that we do not have data on the recall for contributions that are bad quality, but are not short-lived. We focused on short-lived contributions, since we assumed that any truly poor-quality contribution would be undone in short order. It would be of interest to study the relationship between author reputation, and the quality of their contributions (short-lived or not) as perceived by human subjects.

The fact that so few of the short-lived edits performed by high-reputation authors were judged to be of bad quality points to the fact that edits can be undone for reasons unrelated to quality. Many Wikipedia articles deal with current events; edits to those articles are undone regularly, even though they may be of good quality. Our algorithms do not treat in any special way current-events pages. Other Wikipedia edits are administrative in nature, flagging pages that need work or formatting; when these flags are removed, we classify it as text deletion. Furthermore, our algorithms do not track text across articles, so that when text is moved from one article to another, they classify it as deleted from the source article.

From Table 1, we note that the precision is low, by search standards. Our problem, however, is a prediction problem, not a retrieval problem, and thus it is intrinsically different. The group of authors with low reputation includes many authors who are who are good

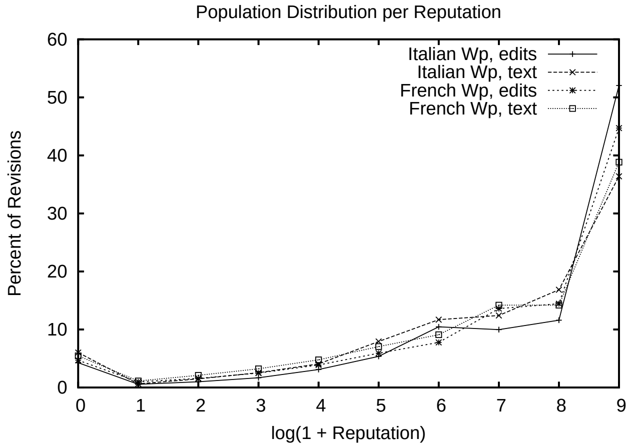

Figure 2: Percentage of text and edit that originated from authors of a certain reputation, in the Italian and French Wikipedias.

contributors, but who are new to the Wikipedia, so that they have not had time yet to build up their reputation.

Figure 2 provides a breakdown of the amount of edits and text additions performed, according to the reputation of the author, for the French and Italian Wikipedias.

Table 3 provides a more in-depth look at the relationship between author reputation and edit longevity on the French Wikipedia. The table shows that, the higher an author's reputation, the higher the longevity of the edits is. Specifically, as author reputation increases, the percentage of edits that have longevities below specified thresholds decreases. For instance, while 25.63% of the edits performed by authors in the lowest reputation bin ($\log(1 + \mathit{rep}) < 1$) have $\alpha_e < -0.8$, only 3.17% of the edits performed by authors in the highest reputation bin ($\log(1 + \mathit{rep}) \geq 9$) do. The corresponding data for the Italian Wikipedia is given in Table 4.

Tables 5 and 6 provide another look at the data, by plotting how the edits in each longevity range were distributed according to author reputation. Each column includes longevity values from the number heading the column, to the number immediately to its right: for instance, the column labeled "0.4" corresponds to the longevity range $\alpha_e \in [0.4, 0.6)$. The rightmost column corresponds to the edit longevity value $\alpha_e = -1$. From Table 5 we see that fully 47.74% of the edits which were immediately reversed ($\alpha_e = -1$) were performed by authors of reputation in the lowest bin. Many columns (see, e.g., those for $\alpha_e = -1$ or $\alpha_e = -0.8$) show an increase in the number of short-longevity edits, as the reputation increases past the first bin. This is not due to the fact that authors of greater reputation are more likely to produce short-longevity edits; indeed, Table 3 showed the opposite to be true. The proportion of short-longevity edits, for bins greater than the first, increases with the reputation, because the number of such edits increases even faster (see column "%data"). For instance, 2.68% of the edits with $\alpha_e = -0.6$ fall in reputation bin 2, which accounts for 1.55% of the edits, while 36.19% of the edits with $\alpha_e = -0.6$ fall in reputation bin 9, which accounts for 45.39% of the edits. This means that authors in reputation bin 2 are $(2.68/1.55)/(36.19/45.39) = 2.17$ times as likely to perform an edit with $\alpha_e = -0.6$ than authors in reputation bin 9.

The corresponding tables for text longevity are Tables 7, 8, 9, and 10, and they tell a very similar story.

5.1 Comparison with Edit-Count Reputation

We compared the performance of our content-driven reputation to another basic form of reputation: *edit count.* It is commonly believed that, as Wikipedia authors gain experience (through revision comments, talk pages, and reading articles on Wikipedia standards), the quality of their submissions goes up. Hence, it is reasonable to take edit count, that is, the number of edits performed, as a form of reputation. We compare the performance of edit count, and of content-driven reputation, in Table 11. As we can see, according to our metrics, content-driven reputation performs slightly better than edit-count reputation on both the Italian and French Wikipedias.

We believe that one reason edit-count based reputation performs well in our measurements is that authors, after performing edits that are often criticized and reverted, commonly either give up their identity in favor of a "fresh" one, thus zeroing their edit-count reputation (thus "punishing" themselves), or stop contributing to the Wikipedia. However, we believe that the good performance of edit count is an artifact, due to the fact that edit count is applied to an already-existing history of contributions. Were it announced that edit count is the chosen notion of reputation, authors would most likely modify their behavior in a way that both rendered edit count useless, and damaged the Wikipedia. For instance, it is likely that, were edit count the measure of reputation, authors would adopt strategies (and automated robots) for performing very many unneeded edits to the Wikipedia, causing instability and damage. In other words, edit count as reputation measure has very little prescriptive value. In contrast, we believe our content-driven reputation, by prizing long-lasting edits and content, would encourage constructive behavior on the part of the authors.

		Cumulative Distribution of Content-Driven Reputation over Edit Longevity										
Rep	%data	≤ 1.0	≤ 0.8	≤ 0.6	≤ 0.4	≤ 0.2	≤ 0.0	≤−0.2	≤−0.4	≤−0.6	≤−0.8	=−1.0
0.0	6.80	100	34.87	31.51	29.85	28.99	28.47	27.81	27.34	26.43	25.63	16.19
1.0	0.86	100	20.42	17.00	14.94	13.81	13.30	12.68	12.16	11.21	10.35	4.21
2.0	1.55	100	20.18	16.63	14.35	13.31	12.72	11.91	11.19	10.43	9.66	3.67
3.0	2.47	100	17.44	13.28	11.31	10.26	9.58	8.82	8.14	7.17	6.65	2.92
4.0	3.66	100	16.30	12.27	10.05	8.91	8.14	7.45	6.84	6.00	5.45	1.84
5.0	6.06	100	15.09	11.46	9.85	9.00	8.32	7.77	7.28	6.54	6.14	1.92
6.0	7.68	100	14.35	10.81	8.85	7.82	7.06	6.44	5.91	5.06	4.58	1.85
7.0	11.53	100	14.69	11.18	9.34	8.36	7.71	7.04	6.42	5.48	5.02	1.33
8.0	14.01	100	13.51	10.22	8.49	7.52	6.88	6.22	5.61	4.71	4.25	1.32
9.0	45.39	100	12.04	8.89	7.27	6.24	5.60	4.91	4.39	3.52	3.17	0.83

Table 3: French Wikipedia: Cumulative distribution of revisions over edit longevity, grouped by reputation. The Rep column is the scaled reputation $\lfloor \log(1 + rep) \rfloor$, where *rep* is the author reputation. Each row shows how much of the total editing was done by authors of each reputation range (the "%data" column), and how those edits were distributed with respect to edit longevity.

		Cumulative Distribution of Content-Driven Reputation over Edit Longevity										
Rep	%data	≤1.0	≤0.8	≤0.6	≤0.4	≤0.2	≤0.0	≤−0.2	≤−0.4	≤−0.6	≤−0.8	=−1.0
0.0	4.27	100	38.94	36.22	23.09	22.31	22.11	21.29	21.00	14.12	13.82	6.00
1.0	0.54	100	25.94	21.09	19.64	18.93	18.62	18.09	17.90	17.08	16.78	5.63
2.0	0.99	100	16.08	11.63	8.68	7.87	7.26	6.85	5.34	3.78	3.33	0.85
3.0	1.67	100	15.11	10.52	8.74	8.01	7.31	6.28	5.15	4.28	3.87	1.54
4.0	3.09	100	11.94	8.33	6.53	5.85	4.91	4.58	4.07	2.45	2.09	0.77
5.0	5.36	100	11.85	8.94	7.29	6.60	5.99	5.46	4.62	2.23	2.02	0.53
6.0	10.46	100	9.85	7.08	6.12	5.73	5.33	4.70	4.24	3.10	2.43	0.46
7.0	9.97	100	11.29	8.18	6.26	5.08	4.38	3.88	3.21	1.92	1.22	0.39
8.0	11.59	100	15.10	11.15	7.37	6.50	5.90	5.31	4.94	3.93	3.50	0.82
9.0	52.05	100	12.84	10.00	8.27	7.38	6.90	6.37	5.79	4.81	3.42	0.53

Table 4: Italian Wikipedia: cumulative distribution of revisions over edit longevity, grouped by reputation. The Rep column is the scaled reputation $\lfloor \log(1 + rep) \rfloor$, where *rep* is the author reputation. Each row shows how much of the total editing was done by authors of each reputation range (the "%data" column), and how those edits were distributed with respect to edit longevity.

		Distribution of Edit Longevity over Content-Driven-Reputation										
Rep	%data	1.0	0.8	0.6	0.4	0.2	0.0	−0.2	−0.4	−0.6	−0.8	−1.0
0.0	6.80	5.21	6.80	6.52	5.80	5.44	6.74	5.88	7.03	12.18	19.01	47.74
1.0	0.86	0.80	0.87	1.01	0.97	0.67	0.80	0.81	0.92	1.66	1.55	1.57
2.0	1.55	1.45	1.64	2.03	1.60	1.43	1.86	2.02	1.34	2.68	2.74	2.46
3.0	2.47	2.40	3.07	2.80	2.60	2.57	2.82	3.05	2.72	2.86	2.73	3.13
4.0	3.66	3.60	4.40	4.70	4.16	4.37	3.79	4.05	3.51	4.44	3.92	2.92
5.0	6.06	6.05	6.57	5.61	5.19	6.28	4.98	5.42	5.12	5.50	7.57	5.04
6.0	7.68	7.74	8.11	8.67	7.94	8.93	7.18	7.37	7.45	8.31	6.22	6.17
7.0	11.53	11.57	12.09	12.21	11.32	11.50	11.54	12.95	12.47	11.70	12.61	6.67
8.0	14.01	14.25	13.77	13.92	13.58	13.89	13.79	15.61	14.42	14.48	12.15	8.03
9.0	45.39	46.95	42.67	42.53	46.86	44.93	46.51	42.84	45.02	36.19	31.49	16.28

Table 5: French Wikipedia: distribution of revisions over reputation, grouped by edit longevity. The Rep column is the scaled reputation $\lfloor \log(1 + \mathit{rep}) \rfloor$, where *rep* is the author reputation. The first column ("%data") shows the amount of editing that was performed by authors in each reputation range. The other columns show, as percentages, how revisions of a specific edit longevity fall into reputation ranges, and therefore add to 100%.

		Distribution of Edit Longevity over Reputation										
Rep	%data	1.0	0.8	0.6	0.4	0.2	0.0	−0.2	−0.4	−0.6	−0.8	−1.0
0.0	4.27	3.03	3.79	23.33	3.96	1.68	6.29	2.14	21.30	1.32	12.48	30.78
1.0	0.54	0.46	0.85	0.33	0.46	0.32	0.51	0.18	0.32	0.17	2.24	3.64
2.0	0.99	0.97	1.44	1.22	0.96	1.15	0.74	2.66	1.12	0.47	0.92	1.01
3.0	1.67	1.64	2.50	1.24	1.45	2.23	3.08	3.34	1.04	0.72	1.45	3.09
4.0	3.09	3.15	3.63	2.32	2.51	5.49	1.86	2.79	3.62	1.15	1.52	2.86
5.0	5.36	5.49	5.09	3.68	4.42	6.24	5.09	7.98	9.29	1.16	2.98	3.44
6.0	10.46	10.94	9.47	4.17	4.86	7.93	11.97	8.49	8.61	7.30	7.68	5.80
7.0	9.97	10.26	10.13	7.93	14.20	13.22	8.95	11.91	9.26	7.28	3.09	4.64
8.0	11.59	11.42	14.94	18.19	12.12	13.37	12.22	7.61	8.47	5.18	11.61	11.34
9.0	52.05	52.64	48.17	37.61	55.06	48.37	49.30	52.91	36.96	75.26	56.03	33.43

Table 6: Italian Wikipedia: distribution of revisions over reputation, grouped by edit longevity. The Rep column is the scaled reputation $\lfloor \log(1 + \mathit{rep}) \rfloor$, where *rep* is the author reputation. The first column ("%data") shows the amount of editing that was performed by authors in each reputation range. The other columns show, as percentages, how revisions of a specific edit longevity fall into reputation ranges, and therefore add to 100%.

		Cumulative Distribution over Text Longevity					
Rep	%data	≤ 1.0	≤ 0.8	≤ 0.6	≤ 0.4	≤ 0.2	≤ 0.0
0.00	7.30	100	14.00	9.71	6.83	6.36	4.10
1.00	1.08	100	8.01	4.23	3.10	2.40	1.85
2.00	1.97	100	8.70	4.37	2.55	2.30	1.26
3.00	3.02	100	7.49	3.84	2.30	2.05	1.38
4.00	4.51	100	6.55	3.17	1.64	1.37	0.81
5.00	6.35	100	5.49	2.43	1.33	1.19	1.05
6.00	9.30	100	4.62	1.95	1.16	1.04	0.67
7.00	11.68	100	5.43	1.76	0.88	0.75	0.58
8.00	14.32	100	4.06	1.55	0.78	0.65	0.52
9.00	40.47	100	3.96	1.67	0.83	0.70	0.58

Table 7: French Wikipedia: Cumulative distribution of revisions over text longevity, grouped by reputation. The Rep column is the scaled reputation $\lfloor \log(1 + rep) \rfloor$, where *rep* is the author reputation. Each row shows how much of the total text was contributed by authors of each reputation range (the "%data" column), and how the text was distributed with respect to text longevity.

		Cumulative Distribution over Text Longevity					
Rep	%data	≤ 1.0	≤ 0.8	≤ 0.6	≤ 0.4	≤ 0.2	≤ 0.0
0.00	6.01	100	11.48	5.23	4.33	4.23	3.17
1.00	0.63	100	5.70	2.03	1.25	1.18	0.93
2.00	1.49	100	7.82	2.39	1.22	1.13	1.03
3.00	2.55	100	5.32	1.78	1.04	0.51	0.45
4.00	4.07	100	7.27	1.73	0.98	0.75	0.44
5.00	7.92	100	7.14	1.26	0.55	0.44	0.34
6.00	11.69	100	8.44	1.06	0.41	0.36	0.24
7.00	12.40	100	5.81	1.62	0.56	0.42	0.26
8.00	16.85	100	6.77	1.34	0.47	0.39	0.29
9.00	36.40	100	10.81	1.81	0.58	0.44	0.34

Table 8: Italian Wikipedia: cumulative distribution of revisions over text longevity, grouped by reputation. The Rep column is the scaled reputation $\lfloor \log(1 + rep) \rfloor$, where *rep* is the author reputation. Each row shows how much of the total text was contributed by authors of each reputation range (the "%data" column), and how the text was distributed with respect to text longevity.

		Distribution of Text Longevity over Reputation					
Rep	%data	1.0	0.8	0.6	0.4	0.2	0.0
0.00	7.30	6.64	10.97	19.60	20.25	44.85	32.22
1.00	1.08	1.05	1.43	1.14	4.41	1.62	2.15
2.00	1.97	1.90	2.99	3.33	2.94	5.57	2.67
3.00	3.02	2.95	3.86	4.36	4.30	5.50	4.50
4.00	4.51	4.45	5.33	6.45	6.93	6.97	3.91
5.00	6.35	6.35	6.79	6.56	4.96	2.44	7.20
6.00	9.30	9.38	8.68	6.88	6.37	9.45	6.67
7.00	11.68	11.68	15.00	9.52	9.12	5.56	7.23
8.00	14.32	14.52	12.55	10.29	11.03	5.12	8.02
9.00	40.47	41.09	32.40	31.88	29.70	12.92	25.44

Table 9: French Wikipedia: distribution of revisions over reputation, grouped by text longevity. The Rep column is the scaled reputation $\lfloor \log(1 + rep) \rfloor$, where *rep* is the author reputation. Columns show, as percentages, how revisions of a specific text longevity fall into reputation ranges, and therefore add to 100%.

		Distribution of Text Longevity over Reputation					
Rep	%data	1.0	0.8	0.6	0.4	0.2	0.0
0.00	6.01	5.82	5.50	5.48	4.79	36.49	37.97
1.00	0.63	0.65	0.34	0.50	0.35	0.89	1.18
2.00	1.49	1.50	1.18	1.78	1.01	0.82	3.07
3.00	2.55	2.65	1.32	1.92	10.49	0.91	2.30
4.00	4.07	4.13	3.30	3.08	7.22	7.37	3.55
5.00	7.92	8.04	6.81	5.77	6.81	4.56	5.32
6.00	11.69	11.71	12.61	7.73	5.25	7.99	5.50
7.00	12.40	12.78	7.60	13.44	13.21	11.44	6.45
8.00	16.85	17.19	13.41	14.81	10.30	9.64	9.85
9.00	36.40	35.52	47.93	45.48	40.58	19.89	24.82

Table 10: Italian Wikipedia: distribution of revisions over reputation, grouped by text longevity. The Rep column is the scaled reputation $\lfloor \log(1 + \mathit{rep}) \rfloor$, where *rep* is the author reputation. Columns show, as percentages, how revisions of a specific text longevity fall into reputation ranges, and therefore add to 100%.

	Precision		Recall		Boost		Coeff. of constr.	
	Edit	Text	Edit	Text	Edit	Text	Edit	Text
	prec_e	prec_t	rec_e	rec_t	boost_e	boost_t	κ_e	κ_t
Italian Wikipedia:								
Content-driven reputation	14.15	3.94	19.39	38.69	4.03	5.83	3.35	7.17
Edit count as reputation	11.50	3.32	19.09	39.52	3.27	4.91	2.53	6.35
French Wikipedia:								
Content-driven reputation	23.92	5.85	32.24	37.80	4.21	4.51	7.33	6.29
Edit count as reputation	21.62	5.63	28.30	37.92	3.81	4.34	5.61	6.08

Table 11: Summary of the performance of content-driven reputation over the Italian and French Wikipedias. All data are expressed as percentages.

5.2 Text Age and Author Reputation as Trust Criteria

The age of text in the Wikipedia is often considered as an indicator of text trustworthiness, the idea being that text that has been part of an article for a longer time has been vetted by more contributors, and thus, it is more likely to be correct [Cro06]. We were interested in testing the hypothesis that author reputation, in addition to text age, can be a useful indicator of trustworthiness, especially for text that has just been added to a page, and thus that has not yet been vetted by other contributors. Let *fresh text* be the text that has just been inserted in a Wikipedia article. We considered all text that is fresh in all the Italian Wikipedia, and we measured that 3.87 % of this fresh text is deleted in the next revision. In other words, Pr(deleted | fresh) = 0.0387. We then repeated the measurement for text that is both fresh, and is due to a low-reputation author: 6.36 % of it was deleted in the next revision, or Pr(deleted | fresh and low-reputation) = 0.0636. This indicates that author reputation is a useful factor in predicting the survival probability of fresh text, if not directly its trustworthiness. Indeed, as remarked above, since text can be deleted for a number of reasons aside from bad quality, author reputation is most likely a better indicator of trustworthiness than these figures indicate.

6 Conclusions

After comparing edit and text longevity values with user quality ratings for revisions, we believe that the largest residual source of error in our content-driven reputation lies in the fact that our text analysis does not include specific knowledge of the Wikipedia markup language and Wikipedia conventions. We plan to make the text analysis more precise in future work.

It is occasionally claimed that the reputation of Wikipedia authors should be domain-specific, so that a good reputation for writing about mathematics does not automatically translate to a good reputation for writing about history. We are not sure domain-specific reputation has more prediction value than generic reputation, such as the one we build in this paper. We suspect that a sense of one's own abilities and limitations, and a knowledge of Wikipedia customs, may play a more important role than domain-specific knowledge in determining the average quality of an author's contributions. The techniques developed for this paper enable us to easily measure the contributions of authors to various categories of the Wikipedia, so we plan to study the usefulness of domain-specific reputation from an experimental point of view in future work.

Acknowledgements We thank Marco Faella for helping organize the manual rating of revisions to the Italian Wikipedia.

References

[BL97] R.C. Burns and D.D.E. Long. A linear time, constant space differencing algorithm. In *Performance, Computing, and Communication Conference (IPCCC)*, pages 429–436. IEEE International, 1997.

[CAM02] G. Cobéna, S. Abiteboul, and A. Marian. Detecting changes in XML documents. In *Proc. of the 18th Intl. Conf. on Data Engineering (ICDE).* IEEE Computer Society Press, 2002.

[CGM97] S. Chawate and H. Garcia-Molina. Meaningful change detection in structured data. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 26–37. ACM Press, 1997.

[CL01] W. Cunningham and B. Leuf. *The Wiki Way. Quick Collaboration on the Web.* Addison-Wesley, 2001.

[Cro06] T. Cross. Puppy smoothies: Improving the reliability of open, collaborative wikis. *First Monday*, 11(9), September 2006.

[CT91] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* J. Wiley & Sons, 1991.

[Del03] C. Dellarocas. The digitization of word-of-mouth: Promises and challenges of online reputation systems. *Management Science*, October 2003.

[Gil05] J. Giles. Internet encyclopaedias go head to head. *Nature*, pages 900–901, December 2005.

[GKRT04] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proc. of the 13th Intl. Conf. on World Wide Web*, pages 403–412. ACM Press, 2004.

[Gol05] J.A. Golbeck. *Computing and Applying Trust in Web-Based Social Networks.* PhD thesis, University of Maryland, 2005.

[Hir77] D.S. Hirschberg. Algorithms for the longest common subsequence problem. *J. ACM*, 24(4):664–675, 1977.

[HM75] J.W. Hunt and M.D. McIlroy. An algorithm for differential file comparison. Computer Science Technical Report 41, Bell Laboratories, 1975.

[HR06] M. Hickman and G. Roberts. Wikipedia — separating fact from fiction. *The New Zealand Herald*, Feb. 13 2006.

[Kle99] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[KSGM03] S.D. Kamvar, M.T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proc. of the 12th Intl. Conf. on World Wide Web*, pages 640–651. ACM Press, 2003.

[LRBPR99] D. Lucking-Reiley, D. Bryan, N. Prasad, and D. Reeves. Pennies from Ebay: The determinants of price in online auctions. Working paper, Vanderbilt University, 1999.

[LZ05] V.B. Livshits and T. Zimmerman. Dynamine: Finding common error patterns by mining software revision histories. In *ESEC/FSE*, pages 296–305, 2005.

[Mye86] E.W. Myers. An o(ND) difference algorithm and its variations. *Algorithmica*, 1(2):251–266, 1986.

[PBMW98] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[RZFK00] P. Resnick, R. Zeckhauser, E. Friedman, and K. Kiwabara. Reputation systems. *Comm. ACM*, 43(12):45–48, 2000.

[Str06] R. Stross. Anonymous source is not the same as open source. *The New York Times*, Mar. 12 2006.

[Tic84] W.F. Tichy. The string-to-string correction problem with block move. *ACM Transactions on Computer Systems*, 2(4), 1984.

[VWD04] F. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 575–582, 2004.

[ZAD$^+$06] H. Zeng, M.A. Alhoussaini, L. Ding, R. Fikes, and D.L. McGuinness. Computing trust from revision history. In *Intl. Conf. on Privacy, Security and Trust*, 2006.

[ZS89] K. Zhang and D. Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, 18(6), 1989.