QuWi: Quality Control in Wikipedia

Alberto Cusinato
Department of Mathematics
and Computer Science
University of Udine
Udine, Italy
alberto.cusinato@
gmail.com

Vincenzo Della Mea
Department of Mathematics
and Computer Science
University of Udine
Udine, Italy
dellamea@dimi.uniud.it

Francesco Di Salvatore
Department of Mathematics
and Computer Science
University of Udine
Udine, Italy
francesco.disalvatore@
gmail.com

Stefano Mizzaro
Department of Mathematics
and Computer Science
University of Udine
Udine, Italy
mizzaro@dimi.uniud.it

ABSTRACT

We propose and evaluate QuWi (Quality in Wikipedia), a framework for quality control in Wikipedia. We build upon a previous proposal by Mizzaro [11], who proposed a method for substituting and/or complementing peer review in scholarly publishing. Since articles in Wikipedia are never finished, and their authors change continuously, we define a modified algorithm that takes into account the different domain, with particular attention to the fact that authors contribute identifiable pieces of information that can be further modified by other authors.

The algorithm assigns quality scores to articles and contributors. The scores assigned to articles can be used, e.g., to let the reader understand how reliable are the articles he or she is looking at, or to help contributors in identifying low quality articles to be enhanced. The scores assigned to users measure the average quality of their contributions to Wikipedia and can be used, e.g., for conflict resolution policies based on the quality of involved users.

Our proposed algorithm is experimentally evaluated by analyzing the obtained quality scores on articles for deletion and featured articles, also on six temporal Wikipedia snapshots. Preliminary results demonstrate that the proposed algorithm seems to appropriately identify high and low quality articles, and that high quality authors produce more long-lived contributions than low quality authors.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Citation analysis, social networks for information retrieval*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WICOW09, April 20, 2009, Madrid, Spain.
Copyright 200X ACM 978-1-60558-488-1/09/04 ...$5.00.

General Terms

Measurement

Keywords

Wikipedia, Quality Control, QuWi, Reputation

1. INTRODUCTION

Quality is an important issue in Wikipedia. The understanding of how Wikipedia works immediately raises obvious concerns about reliability of the contents, trust and reputation of the contributors, and in general credibility of the whole system. Stating it bluntly, if everyone can edit Wikipedia pages, how can their quality be kept to a reasonable level?

As the amount of articles and revisions on Wikipedia is huge, it is practically unfeasible to introduce any quality evaluation system based on classical expert review. Therefore, quality has been taken into account within Wikipedia by means of several approaches. Editors can guide the collaborative editing and writing process; articles can be proposed for deletion; featured articles, i.e., high quality articles, are proposed, nominated and emphasized. The quality issue has also recently gained importance and attention from a wide audience through the famous analysis published in Nature [7]. Finally, several researchers work on alternative quality models.

In this paper we propose and evaluate QuWi, a framework for quality control in Wikipedia. We build upon a previous proposal by Mizzaro [11], who proposed a method for substituting and/or complementing peer review in scholarly publishing. Since the Wikipedia scenario is different from the scholarly publishing (articles are never finished, and their authors change continuously), we adapt Mizzaro's model: we define a modified algorithm that takes into account the different domain, with particular attention to the fact that authors contribute identifiable pieces of information that can be further modified by other authors.

This paper is organized as follows. In Section 2 we briefly survey the approaches to quality in Wikipedia and we summarize Mizzaro's model. In Section 3 we describe our approach, named QuWi, by emphasizing the changes to Mizzaro's peer review model. Section 4 reports on an experimental evaluation carried on by applying the QuWi obtained model to a subset of Wikipedia articles with their complete evolution. Section 5 closes the paper.

2. RELATED WORK

The background for this paper is divided into two parts: previous practices and research concerning quality in Wikipedia and Mizzaro's model for quality in scholarly publishing.

2.1 Quality in Wikipedia

Wikipedia developed a number of internal norms and habits aimed at promoting quality: these include the already mentioned featured articles and article deletion policy. In particular, the acceptance process for a featured articles has been recognized as selecting good quality articles [13, 14]. As an alternative to deletion, an Editor may decide to attract contributors attention towards low quality articles by means of specific tags, aimed at explicitly asking for enhancements. Also some automatic enhancement is provided through the so called *bots*, i.e., software programs able to navigate Wikipedia, delete spam, correct small errors, etc.

A large project like Wikipedia, with an absolutely liberal active participation policy, yet with interesting practical results, attracted research on quality issues from at least two different points of view: its evaluation, and methods to enhance it.

Concerning evaluation, in 2005 a paper has been published on Nature [7] that compared 42 scientific terms in Wikipedia and Encyclopedia Britannica. Major and minor errors were found in both sources, without a significantly higher quality of the more traditional encyclopedia. Lih [10] examined the number of authors and modifications of a set of high quality articles and demonstrated that the more the authors, the higher the article quality. Another positive result comes from a work of Emigh and Herring [6], that proposed an evaluation based on language formality, and compared Wikipedia with the Columbia encyclopedia and the online community Everything2 (`http://everything2.com/`). Briefly, the language used in Wikipedia is more similar to the one used in the traditional encyclopedia than the one coming from the online community, which by the way also provides for an experience-based reputation score. Stvilia et al. realized a framework for Information Quality Assessment [12], that applied to Wikipedia helped in defining critical activities, main problems, and their probabilities. Evaluation metrics obtained from that effort allowed to recognize an higher quality in featured articles than in randomly chosen articles.

Other research is aimed at providing tools for actively helping in enhancing quality. A concept related to quality is *reputation*, seen as the recognizable quality of contributors. Anthony, Smith, and Williamson [3] defined as *zealots* the registered users that make many contributions, while anonymous users with few contributions were called *Good Samaritans*: both provide for good quality contributions, with slightly different patterns. In fact, zealots contribution quality grows with the number of contributions, while good samaritans contributions quality is higher when they contribute rarely. On the other side, anonymous users with many contributions are not interested in reputation, and often are responsible for vandalisms and spam.

Another way for measuring quality is to adapt external popularity measures like PageRank [5] and HITS [9]: both correlate well with article quality [4,15]. An approach based on interaction data between articles and their contributors derived from the article edit history has been also described [8].

Finally, an interesting approach to quality is the one proposed by Adler and De Alfaro [1, 2]. They describe an extension of Wikipedia interface, where text background in articles is colored depending on author's reputation, so that readers may have insights on the article quality at a glance. Reputation is calculated through implicit voting by users: in fact, if an user that contributes to an article does not modify or delete text of another author, this is considered as a positive judgment on that text.

2.2 Quality in scholarly publishing: Mizzaro's model

This section intuitively summarizes the basic idea on which the system proposed by Mizzaro in [11] is based. More details are provided in the original article. The algorithm exploits revision history to compute author's quality in a loosely similar way as it has been exploited by Zeng et al. [16] to compute trust.

Let us imagine a scholarly journal in which each paper is immediately published after its submission, without a refereeing process. Each paper has some scores, measuring its quality (accuracy, comprehensibility, novelty, and so on). For the sake of simplicity, a single score, measuring overall quality, is used but the generalization to multi-dimensional quality measures is straightforward. This score is initially zero, or some predetermined value, and it is later dynamically updated on the basis of the readers' judgments. A subscriber to the journal is an author or a reader (or both). Each subscriber has a score too, initially set to zero (or some predetermined value) and later updated on the basis of the activity of the subscriber (if the subscriber is both an author and a reader, she has two different scores, one as an author and one as a reader). Therefore, the scores of subscribers are dynamic, and change accordingly to subscribers' behavior: if an author with a low score publishes a very good paper, i.e., a paper judged very positively by the readers, her score increases; if a reader expresses an inadequate judgment on a paper, her score decreases accordingly, and so on.

Every object with a score (author, reader, paper) also has a *steadiness* value, representing how much steady the score is: for instance, old papers, i.e., papers that have been read and judged by many readers, will have a high steadiness; new readers and authors will have a low steadiness. Steadiness affects the score update: a low (high) steadiness allows quicker (slower) changes of the corresponding score. While a score changes, the corresponding steadiness value increases.

As time goes on, readers read the papers, judgments are expressed, and the corresponding scores and steadinesses vary accordingly. The score of a paper can be used for deciding to read or not to read that paper; the scores of authors and readers are a measure of their research productivity, then they will try to do their best for keeping their score at a high level, hopefully leading to a virtuous circle (publishing good papers and giving correct judgments to the read papers). A steadiness value is an estimate of how stable and, therefore, reliable the corresponding score is.

For understanding the details of the automatically refereed journal proposed here, let us follow the events that happen when a paper is read and judged by a reader:

1. **Paper.** First of all, the paper score is updated: if the judgment is lower (higher) than the actual paper score, the paper score decreases (increases). The score of the reader determines the weight of the judgment: judgments given by higher-rated readers will be more important, and will lead to higher changes, than judgments given by lower-rated readers.

   The steadiness of the paper increases, since the score of the paper is now computed on the basis of one more judgment, and is therefore statistically more reliable.

2. **Author.** Then, the author's score is updated: when the score of a paper written by an author decreases (increases), the score of the author decreases (increases). Authors' scores are linked to the scores of their papers.

   The steadiness of the author, similarly to the steadiness of the paper, increases, since the score of the author is now obtained with one more judgment and is therefore statistically more reliable.

3. **Reader.** Then the reader's score is updated: if one reader's judgment about a document is "wrong" (i.e., as we will see in the next section, too far from the average), the reader's score has to decrease. Therefore, the reader's score is updated depending on the *goodness* of her judgment, i.e., how much adequate her judgment is, or how much it agrees with the current score of the paper.

   Again, the steadiness of the reader increases, since her score, computed on the basis of the goodness of her judgments, is obtained on the basis of one more judgment.

4. **Previous readers.** Finally, the scores of the readers that previously read the same paper are updated: if a judgment causes a change in a paper score, all the goodnesses of the previously expressed judgments on that paper have to be re-estimated. Therefore, a judgment on a certain paper leads to an updating of the scores of all the previous readers of that paper.

   Again, the steadinesses of the previous readers increase since the goodnesses of the readers, that lead to their scores, are obtained on the basis of one more judgment.

The updating of the scores of the previous readers deserve further explanation. After the paper score has changed, it is possible to revise the goodness of the old readers' judgments, and to update the old readers' score consequently: for instance, if an old reader r expressed a judgment j that was "bad" (distant from the paper score) at that time, but after that the paper score changes and becomes more similar to j, then the score of r (s_r) has to increase. Let us take into account a simple concrete example (Figures 1, 2, and 3, all from [11], show the temporal evolution):

- At time t_0, we have a paper p with score $s_p(t_0)$, three readers r_1, r_2, and r_3 with their scores $s_{r_1}(t_0)$, $s_{r_2}(t_0)$, and $s_{r_3}(t_0)$.

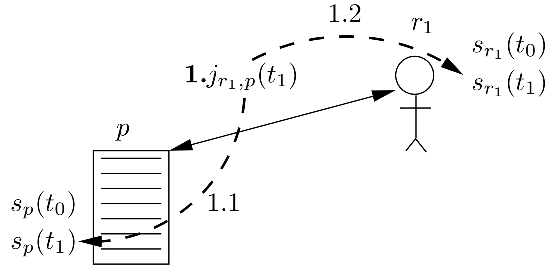

Figure 1: Updating of previous readers' scores: t_1.

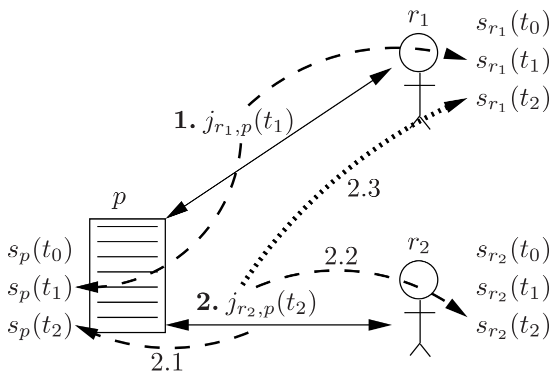

Figure 2: Updating of previous readers' scores: t_2.

- At a following time instant $t_1 > t_0$ (Figure 1), reader r_1 reads paper p expressing the judgment $j_{r_1,p}(t_1)$ (continuous double arrow line in figure). This causes the updating of the scores of p (dashed line in figure labeled with 1.1) and r_1 (dashed line labeled with 1.2), obtaining $s_p(t_1)$ and $s_{r_1}(t_1)$.

- At time $t_2 > t_1$ (Figure 2), reader r_2 reads p expressing $j_{r_2,p}(t_2)$. The scores of p and r_2 are updated consequently, leading to $s_p(t_2)$ and $s_{r_2}(t_2)$ (dashed lines labeled with 2.1 and 2.2). But also the score of r_1 has to be updated (dotted line labeled with 2.3), since the goodness estimated at time t_1 for $j_{r_1,p}(t_1)$ with respect to $s_p(t_1)$ has to be re-estimated now that the score of p is $s_p(t_2)$.

- At time $t_3 > t_2$ (Figure 3), r_3 reads p expressing $j_{r_3,p}(t_3)$. This changes the score of p ($s_p(t_3)$, dashed line labeled with 3.1), the score of r_3 ($s_{r_3}(t_3)$, dashed line labeled with 3.2), and the scores of the previous two readers ($s_{r_2}(t_3)$ and $s_{r_1}(t_3)$, dotted lines labeled with 3.3 and 3.4).

In other words, the goodness of a reader's judgment is an approximation of the *ideal goodness*, defined as the difference between the reader's judgment and the final score of the paper (i.e., the score obtained when the last judgment on that paper has been expressed). Since the final score is obviously not available when the judgment is expressed, it has to be estimated (updating of the reader), but this estimate is revised and refined as time evolves and tends to $+\infty$ (updating of previous readers).

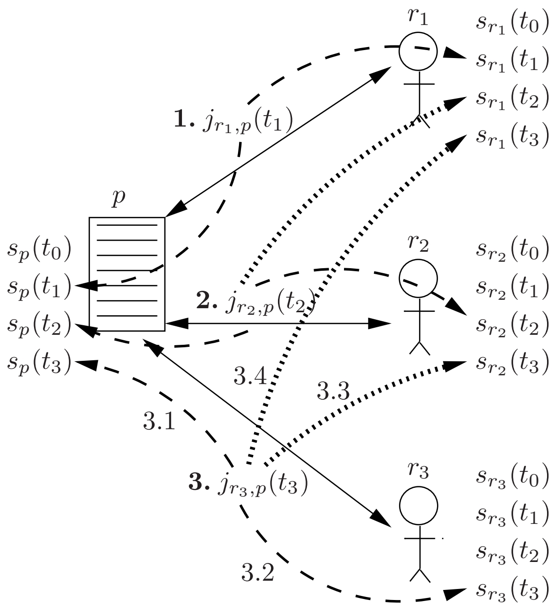

Figure 3: Updating of previous readers' scores: t_3.

In [11] the results of some software simulations demonstrate the effectiveness of the approach, and its resilience to different kinds of malicious behaviors. The aim of the present paper is to adapt Mizzaro's model to the Wikipedia case, and to experimentally evaluate its effectiveness.

3. THE PROPOSED METHOD: QuWi

3.1 Two problems

When trying to adapt Mizzaro's proposal to the Wikipedia case, one has to face two problems. First, in Wikipedia there is no way (currently) for judging an article. Thus, some form of implicit evaluation is needed. Second, in Wikipedia there is no fixed set of authors of an article, as they change over time, with the article evolution. Thus, a method for tracking individual contributions is needed.

The first problem can be dealt with in the following way. The Wiki philosophy suggests a reader to modify an article, if it is believed in some way imprecise. Thus, a reader modifying an article, by means of deletion, insertion, or revert operations, expresses a judgment on the article: in particular, a positive judgment for what she left unmodified, and a negative judgment for what she modified or deleted. Moreover, the smaller (larger) the modification, the more positive (negative) the judgment. This observation allows to define an implicit scoring method like in [2]: the score given by a reader is the ratio between the unmodified text and the original text size (thus, within the [0..1] range).

We should also consider that the article as considered in the original proposal cannot be used to describe the atomic entity underlying Wikipedia publishing system. In fact, each author is responsible for one or more contributions that may range from one character to a whole article: that *contribution* is the entity to which we intend to refer.

To solve the second problem, i.e., tracking individual contributions, a method is needed to establish who is the author of every contribution that generated an article. Wikipedia is currently using a method for calculating differences among versions of the same article, but it is imprecise and unable to recognize some particular cases, typically involving text chunks movement in an article. This urged the development of a more precise algorithm for comparing consecutive versions of an article, to extract single contributions to be associated to their authors. We come back later to this two problems; we introduce some notation first.

3.2 Notation

We define the following notation to identify entities involved in the scoring method:

- t is a discrete time instant, t_{i+1} is the next one. Every instant corresponds to the expression of a score: a score expressed at time t influences values at time t_{i+1}.
- p is a Wikipedia article ("p" is a mnemonic for "paper"; we reserve "a" for "author", see below).
- $s_e(t)$, $s_a(t)$, $s_r(t)$ are, in order, the scores for a contribution ("e" stands for edit), an author, and a reader at time t.
- $s_{we}(t)$, $s_b(t)$, $s_p(t)$ are, in order, scores for survived words in a contribution, for a block and for a page at time t.
- $\sigma_e(t)$, $\sigma_a(t)$, $\sigma_r(t)$ are, in order, steadinesses of a contribution, of an author, and of a reader.
- $\sigma_{aMAX}(t)$ and $\sigma_{rMAX}(t)$ are maximum steadiness as a reader and as an author at time t.
- $t_{r,e}$ is the time point when the reader r expressed her score on e (a reader cannot express more than one score for the same contribution, however she can update her score by deleting the previous one).
- $j_{r,e}(t)$ is the score expressed by reader r at time t on contribution e. For the sake of simplicity, $j_{r,e}(t_{r,e})$ is rewritten as $j_{r,e}$.
- $w_{j_{r,e}}(t)$ is the weight of score j expressed by reader r on contribution e at time t. For the sake of simplicity, $w_{j_{r,e}}(t_{r,e})$ is rewritten as $w_{j_{r,e}}$.
- $g_{j_{r,e}}(t)$ is the goodness at time $t$of judgment j expressed by reader r on contribution e.
- $R_e(t)$ is the set of readers that read e before time t.
- $E_a(t)$ is the set of contributions of author a before time t.
- $E_r(t)$ is the set of contributions read by reader r before time t.
- $W_b(t)$ and $W_b(t)$ are the set of words at time t respectively in block b and in article p.
- d_e is the size (number of characters) of contribution e.

We also remark that in our model, as in the original one in [11], article, reader, and author scores are processed uniformly; therefore in the following we will use the single term *quality* to refer to credibility, reputation, and trust of articles, author, and readers.

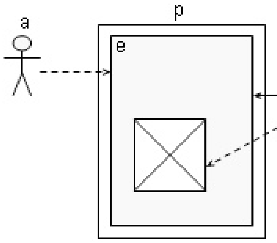

Figure 4: A reader r deleting part of a contribution e by author a in article p.

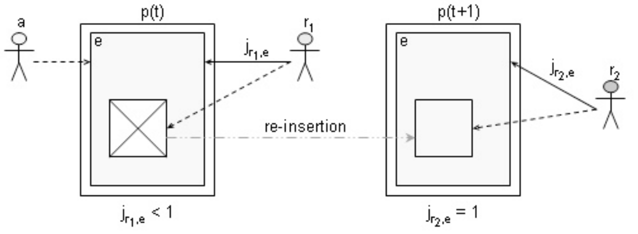

Figure 5: A reader r_2 re-inserting the part of a contribution e by author a in article p that had been deleted by a previous reader r_1.

3.3 Implicit judgments

Since, as remarked above, no explicit judgments are possible in current Wikipedia, we resorted to implicit judgments, inferred by the activities carried on by Wikipedia contributors. Consider for instance Figure 4: the "reader" (actually, she is a contributor) r, by deleting a portion of the text e written by a expresses a judgment $j_{r,e}$ on the contribution e. $j_{r,e}$ is lower as the portion of deleted text becomes larger: the best judgment is given by no deletion; the worst judgment by deleting all the contribution e.

Figure 5 goes on with the example by assuming that, later on, another reader r_2 re-inserts the deleted text (revert operations are one click away in the Wikipedia revision system, since they are handy to remove spam and vandalism). This is interpreted as r_2 giving a low judgment to the contribution by author/reader r_1, and a high judgment to the initial contribution e. Similar implicit judgments have been defined for all kinds of operations (insert, delete, revert, move).

Also, we take into account reader's focus of attention: it seems reasonable to assume that the judgment by a reader is stronger and more reliable on the part of page that is closer to the text being modified; indeed a reader might not read at all the text far away in a long page. Figure 6 shows an example: when a/r_3 adds a contribution e_3, she is expressing a mild positive judgment on previous contribution e_1 (e_3 is assumed to be small) and e_2, which is left unmodified. However, the weight of the judgment on e_2 is lower since a/r_3 attention is on e_1, not on e_2: this is modeled by a weight constant $c_w < 1$.

It should be noted that each user may have both roles of reader and author of the same article, depending on the kind of operations she carries out on the text. The silent user, i.e., the reader that does not become author too, cannot be tracked unless using access logs, which are not currently made available by Wikipedia. Thus, the reader that reads an article and decides not to modify it because already perfect is not considered by this method. However, it is difficult to distinguish such a reader from the readers unable to modify the article for technical difficulties or laziness, and finally it is easy to generate fake HTTP requests to simply access articles, so counting page views is not a viable solution.

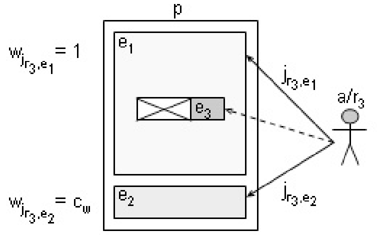

Figure 6: Reader's attention.

3.4 Tracking individual contributions

The main aim of the algorithm for tracking individual contributions is to recognize the basic operations of deletion, insertion, and revert of text chunks between two consecutive revisions of an article. As Wikipedia stores all complete revisions of all articles, the algorithm starts from that to find out the basic operations. To reach that aim, each *article* is divided into *blocks* (i.e., paragraphs), each block into *blockedits* (i.e., chunks of survived text), *dels* (i.e., single chunks of deleted text), and *editdels* (i.e., a deletion operation involving one or more chunk of text). An *edit* is a contribution consisting of new text from a *user*. Carrying out some modification on the article also causes the implicit assignment of a *judgment* to an edit, including a weight simulating author's attention (higher in chunks close to edit, weaker in more distant text).

3.5 The scoring algorithm

The formulas defined in Mizzaro's model have been adapted to the finer granularity needed for Wikipedia. Novel ingredients thus include, e.g., attention and contribution size. Details are as follows.

The score of a contribution e is defined as weighted average of its readers' judgments, where weights are the reader score at the time she expressed her judgement, and reader attention:

$$s_e(t) = \frac{\sum_{r \in R_e(t)} s_r(t_{r,p}) \cdot w_{j_{r,e}} \cdot j_{r,p}}{\sum_{r \in R_e(t)} s_r(t_{r,p}) \cdot w_{j_{r,e}}}.$$

The steadiness of a contribution e is defined as the sum of score of its reader, weighted with attention:

$$\sigma_e(t) = \sum_{r \in R_e(t)} s_r(t_{r,p}) \cdot w_{j_{r,e}}.$$

The score of an author a is defined as the weighted average of her article scores, where weights are steadiness and contribution size:

$$s_a(t) = \frac{\sum_{e \in E_a(t)} \sigma_e(t) \cdot d_e \cdot s_e(t)}{\sum_{e \in E_a(t)} \sigma_e(t) \cdot d_e}.$$

The steadiness of an author a is defined as the weighted average of the steadinesses of her contributions, weighted with contribution sizes:

$$\sigma_a(t) = \sum_{e \in E_a(t)} \sigma_e(t) \cdot d_e.$$

The judgment expressed by reader r on contribution e is defined as the ratio between the size of contribution e after r contribution and the original size of e:

$$j_{r,e}(t) = \frac{|e(t_{r,e})|}{|e(t_{a,e})|}.$$

The goodness of a judgment j at time t is defined as the distance between judgment and the contribution score at time t:

$$g_{j_{r,e}}(t) = 1 - \sqrt{|j_{r,e} - s_e(t)|}.$$

The score of a reader r is defined as the weighted average of her judgment goodness, where weights are given by contribution steadinesses:

$$s_r(t) = \frac{\sum_{e \in E_r(t)} \sigma_e(t) \cdot g_{j_{r,e}}(t)}{\sum_{e \in E_r(t)} \sigma_e(t)}.$$

The steadiness of a reader r is defined as the sum of steadinesses of contributions she judged:

$$\sigma_r(t) = \sum_{e \in E_r(t)} \sigma_e(t).$$

The score of survived words in a contribution is:

$$s_{we}(t) = s_a(t_{a,e}) \cdot \frac{\sigma_a(t_{a,e})}{\sigma_{aMAX}(t_{a,e})} + \sum_{r \in R_e(t)} s_r(t_{r,e}) \cdot \frac{\sigma_r(t_{r,e})}{\sigma_{rMAX}(t_{r,e})}.$$

The block score is given by the average score of its words, weighted by the word size (number of characters):

$$s_b(t) = \frac{\sum_{w \in W_b(t)} |w| \cdot s_w(t)}{\sum_{w \in W_b(t)} |w|}.$$

Finally, the article score is the average score of its words, weighted by the word size:

$$s_p(t) = \frac{\sum_{w \in W_p(t)} |w| \cdot s_w(t)}{\sum_{w \in W_p(t)} |w|}.$$

To update scores for all involved entities, we defined a number of formulas, modified from Mizzaro proposal, that allow to compute a score at time $t+1$ starting from score at time t. Therefore we do not need the system to compute the long sums in the above formulas.

4. EXPERIMENTAL EVALUATION

4.1 Methods and data

In order to evaluate whether the quality values attributed by QuWi to articles and contributors are meaningful, we run some experiments on real data. We downloaded from the Wikipedia site the complete dump of Italian Wikipedia, including article history until June 2007. Then, we extraceted the whole Science category. At that time, the selected category included 19917 articles with their own history, consisting in 482513 revisions that involved 10526 contributors. In addition to initial and final revisions, we also identified five intermediate revisions, equally distributed every 67610 revisions, thus obtaineing six temporal snapshots. We then run the proposed algorithm on the whole set; we recorded the score values on each of the six temporal snapshots.

To compare the obtained scores with some ground truth, we exploited two of the traditional quality control methods available in Wikipedia: we chose the *featured articles* and the *articles proposed for deletion* (we could not use the deleted articles since they are physically removed from the dump) as representative of good and bad articles, respectively. The former are chosen among articles consensually recognized as good, the latter among those that are not good, not adequate for an encyclopedia, or controversial. So, we considered human evaluations as gold standard for identifying respectively "good" and "bad" articles. We then analyzed our automatically computed quality measures on them, together with an amount of other measurements aimed at describing contributors, readers, and articles quality from a statistical point of view. Evolution of articles through the six temporal snapshots has been also considered, as well as survival of contributions.

Using the Italian version allowed us an easy manual inspection of article meaning when needed. Also, as stated in a personal email from the *Wikipedia information team*, "the English Wikipedia dump has actually become something of a myth amongst Wikipedian", and we have not been able to download the Wikipedia dump of the English version.

4.2 Results

4.2.1 Descriptive statistics

The number of edits per contributor is distributed according to a Power Law: only 83 contributors edited at least 500 contributions, only 38 more than 1000, only 3 more than 5000. Anonymous contributors accounted for about 15% of contributions. Also the number of characters per contribution is distributed according to a Power Law, i.e., there are few long contributions, and many short contributions. Both observations are in agreement with many similar observations about phenomena involving networks.

4.2.2 Article evolution

When analyzing how articles are modified during their lifetime by authors contributions, it can be said that each article is converging towards some more or less final version, although something to be changed is still there in the final snapshots.

The average score of articles slightly increases in time, as can be seen for the whole sample (in Figure 7) and for the articles present since the first snapshot (in Figure 8).

The average score of all articles in the sample is 0.42; when considering only articles available since the first snapshot (and thus not including articles created later), the average increases to 0.57. This is consistent with the assumption that article quality increases with time.

4.2.3 Contributions survival

At first, author score has been compared with the survival rate of her contributions, to verify whether the two variables are correlated. Figure 9 shows this relationship. Kendall correlation value is 0.88; it is high as expected: contributors with a high score are those having produced long lived contributions.

We also evaluated where the author score is predictive of contribution survival (Figure 10); for this, we correlated the

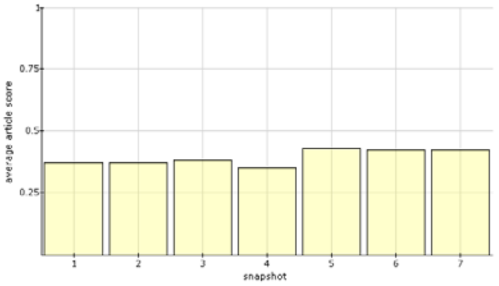

Figure 7: Evolution of average score of all articles

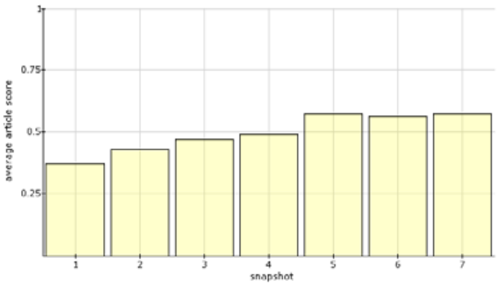

Figure 8: Evolution of average score of articles in Wikipedia since the first snapshot

author initial score with her contribution survival, obtaining a Kendall correlation value of 0.41. Although the value suggests a weak correlation, the figure shows that low scored authors have a larger variability than high scored authors.

On the other side, we also considered how many contributions are fully deleted, depending on author's initial score; results are plotted in Figure 11, and Kendall correlation is 0.53. Variability is similar to that observed for contribution survival.

4.2.4 Bad and good articles

We examined the scores generated by our system in the two particular cases of featured articles and articles for deletion. In the article set we used, 19 featured articles and 75 articles for deletion were present. The average score of featured articles is 0.88 (significantly higher than the 0.42 for the whole sample); it is 0.95 when considering only articles present in Wikipedia since the first snapshot (vs. 0.57 for the whole sample). Figure 12 shows the distribution of scores.

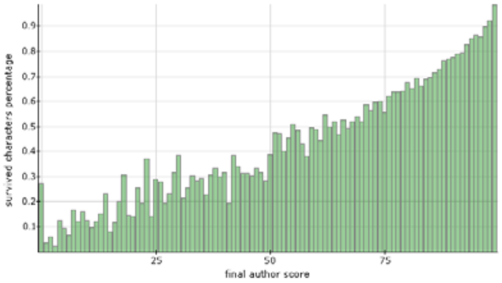

Figure 9: Final author score versus percentage of survived characters

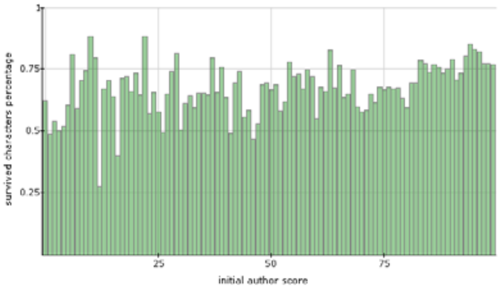

Figure 10: Percentage of survived characters versus author's initial score

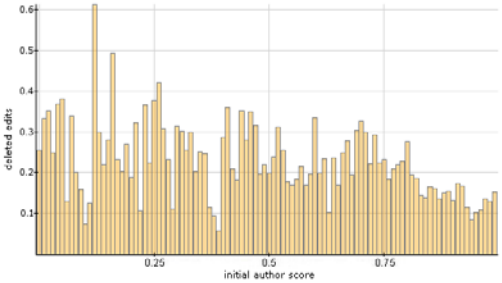

Figure 11: Deleted contributions versus author's initial score

The average score of articles for deletion is 0.27; Figure 13 shows the distribution of scores. Four articles for deletion show a very high score (0.7 or higher): we examined such outliers to understand the reasons for deletion. One article concerns the term "qui" ("here"), that has been considered useless for an encyclopedia, although the description quality was good. For another article, copyright problems were noted at a certain point of its history. Finally, two of those articles were proposed for deletion due to controversial topics ("suicide" and "abortion").

These results clearly show that article scores are consistent with human judgments.

5. DISCUSSION AND CONCLUSIONS

We started from an algorithm originally proposed by Mizzaro for quality control in scholarly publishing [11]. That proposal cannot be directly applied to Wikipedia because scientific articles have a public, final version with a specified number of Authors, while Wikipedia articles are in continuous evolution, with open contribution, and because the original model is based on explicit judgments, which are not considered in Wikipedia. We have adapted the algorithm by taking into account the differences, and evaluated it on a Wikipedia dump. The preliminary experimental results demonstrate that the proposed algorithm QuWi seems to appropriately identify high quality and low quality articles,

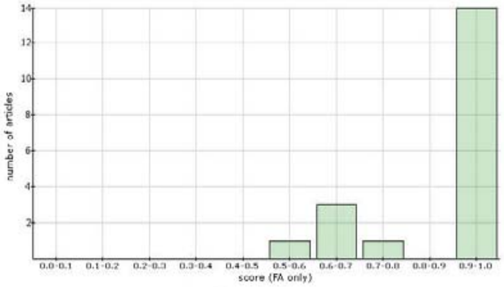

Figure 12: Score distribution of featured articles

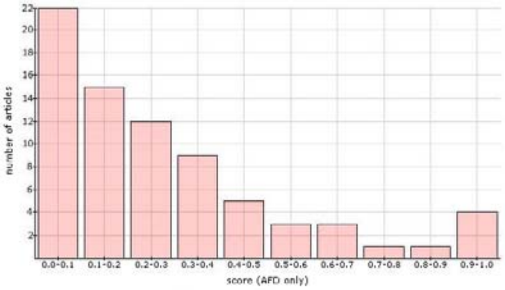

Figure 13: Score distribution of articles for deletion

and that good quality authors produce more long-lived contributions than low quality authors. The number of featured articles involved in the evaluation is low, so that some further investigation is needed, although the results are quite encouraging. While the presented approach, as others, exploits revision history [2, 8], steadiness is a novel element, not found in any of the related studies, and embedded in our model. The algorithm also produces quality information about authors and readers, not presented here for the sake of brevity.

Further work include another run on a wider subset of the English Wikipedia, in order to reach a wider number of featured articles and articles for deletion, thus having more statistically significant data. It would also be interesting to have access to the Wikipedia HTTP logs, to obtain more data to model readers behavior in a more accurate and effective way.

We have also developed (but not yet tested) an experimental interface for coloring contributions basing on quality, similar to the one proposed in [2]. Future user studies will investigate its effectiveness.

6. REFERENCES

[1] B.Thomas Adler, Jason Benterou, Krishnendu Chatterjee, Luca De Alfaro, Ian Pye, and Vishwanath Raman. Assigning Trust to Wikipedia Content. Technical Report UCSC-CRL-07-09, School of Engineering, University of California, Santa Cruz, CA, USA, November 2007.

[2] Thomas B. Adler and Luca De Alfaro. A content-driven reputation system for the Wikipedia. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 261–270, New York, NY, USA, 2007. ACM Press.

[3] Denise Anthony, Sean W. Smith, and Tim Williamson. The Quality of Open Source Production: Zealots and Good Samaritans in the Case of Wikipedia. Technical Report TR2007-606, Dartmouth College, Computer Science, Hanover, NH, September 2007.

[4] Francesco Bellomi and Roberto Bonato. Network Analysis for Wikipedia. *Proceedings of Wikimania*, 2005.

[5] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[6] William Emigh and Susan C. Herring. Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias. *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, 2005.

[7] Jim Giles. Special report: Internet encyclopedias go head to head. *Nature*, 438(15):900–901, 2005.

[8] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. Measuring article quality in wikipedia: models and evaluation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 243–252, New York, NY, USA, 2007. ACM.

[9] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[10] Andrew Lih. Wikipedia as participatory journalism: Reliable source? metrics for evaluating collaborative media as a news resource. In *5th International Symposium on Online Journalism*, 2004.

[11] Stefano Mizzaro. Quality control in scholarly publishing: A new proposal. *Journal of the American Society for Information Science and Technology*, 54(11):989–1005, 2003.

[12] Besiki Stvilia, Les Gasser, Michael B. Twidale, and Linda C. Smith. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12):1720–1733, 2007.

[13] Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. Information quality work organization in Wikipedia. *Journal of the American Society for Information Science*, 59(6):983–1001, 2008.

[14] Fernanda B. Viégas, Martin Wattenberg, and Matthew M. McKeon. The hidden order of wikipedia. *Proceedings of HCII*, 2007, 2007.

[15] Dennis M. Wilkinson and Bernardo A. Huberman. Assessing the Value of Coooperation in Wikipedia. *First Monday 12(4) April*, 2007.

[16] Honglei Zeng, Maher A. Alhossaini, Li Ding, Richard Fikes, and Deborah L. McGuinness. Computing trust from revision history. In *PST '06: Proceedings of the 2006 International Conference on Privacy, Security and Trust*, pages 1–1, New York, NY, USA, 2006. ACM.