Assessing the Quality of Wikipedia Articles with Lifecycle Based Metrics

Thomas Wöhner
Martin-Luther-University Halle-Wittenberg
Universitätsring 3
D-06108 Halle (Saale)
+49 345 55 23478
thomas.woehner@wiwi.uni-halle.de

Ralf Peters
Martin-Luther-University Halle-Wittenberg
Universitätsring 3
D-06108 Halle (Saale)
+49 345 55 23471
ralf.peters@wiwi.uni-halle.de

ABSTRACT
The main feature of the free online-encyclopedia Wikipedia is the wiki-tool, which allows viewers to edit the articles directly in the web browser. As a weakness of this openness for example the possibility of manipulation and vandalism cannot be ruled out, so that the quality of any given Wikipedia article is not guaranteed. Hence the automatic quality assessment has been becoming a high active research field. In this paper we offer new metrics for an efficient quality measurement. The metrics are based on the lifecycles of low and high quality articles, which refer to the changes of the persistent and transient contributions throughout the entire life span.

Categories and Subject Descriptors
H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces – *Computer-supported cooperative work, web-based interaction;* K.4.3 **[Computers and Society]**: Organizational Impacts—*Computer-supported collaborative work*

General Terms
Measurement, Reliability, Experimentation

Keywords
Wikipedia, quality assessment, Wikipedia lifecycle, transient contribution, persistent contribution

1. INTRODUCTION
Web2.0, which is characterized by user-generated content, has been becoming increasingly import in the World Wide Web since the collapse of the "dot-com bubble" in 2001 [14]. The most popular Web2.0 application is the free online-encyclopedia, Wikipedia, which contains more than *10,000,000* articles in over *260* languages, as measured in *October 2008*. The English Wikipedia, which consists of about *2,550,000* articles, is the largest one, followed by the German Wikipedia with about *800,000* articles [23]. According to *Alexa*[1], Wikipedia is constantly listed in the top ten most visited websites worldwide. The main feature of the website is the wiki-tool, which allows viewers to edit the articles directly within the web browser [5]. With Wikipedia the articles are contributed voluntarily by everyday web users, whereas with traditional encyclopedias, the articles are written by experts.

WikiSym '09, October 25-27, 2009, Orlando, Florida, U.S.A.

The openness of the system attracts many volunteers, who write, update and maintain the articles. According to a study of the scientific magazine, *Nature*, the quality of Wikipedia is comparable to that of the traditional Encyclopedia Britannica [7]. On the other hand, the open access has been known to cause quality problems. The possibility of manipulation and vandalism cannot be ruled out. For example, inaccurate information is occasionally published by opportunistic or inexperienced contributors. Additionally, when articles are not being focused on by the Wikipedia community and hence there is a lack of volunteers providing content, these articles can be incomplete or insufficient. As a consequence, the quality and accuracy of any given Wikipedia article cannot be guaranteed.

To overcome this weakness Wikipedia has developed several user-driven approaches for evaluating the articles. High quality articles can be marked as "*Good Articles*" or "*Featured Articles*" whereas poor quality articles can be marked as "*Articles for Deletion*" [22]. However, these user-driven evaluations can only partially solve the problem of quality transparency since only a very small part of Wikipedia is evaluated by them. For example in *January 2008* only about *3,500* of *650,000* articles altogether were evaluated in the German Wikipedia. Another difficulty of the user-driven evaluations is that the Wikipedia content is by its nature highly dynamic and the evaluations often become obsolete rather quickly.

Due to these conditions, recent research work involves automatic quality assessment that is being developed specifically for Wikipedia. In this paper we provide a new approach to using metrics to effectively measure the quality of the Wikipedia articles. The metrics are based on tracking the changes in the editing intensity throughout the entire existence of an article, which we refer to as the lifecycle of the article. In order to calculate the editing intensity, we have constructed two new metrics, which we call the "*persistent contribution*" and the "*transient contribution*". The transient contribution refers to the number of words which were changed and reversed in the same given period of time. Since versions which have been vandalized

[1] www.alexa.com

or versions that include obvious inaccurate information are reverted back in a short period of time [20, 21], such contributions are particularly well covered by the transient contributions. The persistent contribution refers to all effective edits which remain in the article beyond the period.

By using these metrics we are able to construct lifecycles that represent either low or high quality articles. Characteristic differences between these lifecycles can be seen quite clearly, thus providing a useful basis for measuring the quality of any given article. Our analysis reveals that such lifecycle based metrics are more efficient than the word count of an article, which, according to Blumenstock [3], is the most efficient metric currently known.

This paper is structured as follows. First we will describe the related work and explain how our metrics can contribute to the approaches currently being applied for quality assessment. Secondly we will introduce the quality evaluations currently being used with the German Wikipedia, since this has been our test base for identifying high and low quality articles. Thirdly, we will introduce the model we have been using to analyze the Wikipedia articles and illustrate how we compute the transient and persistent contributions based on this model. Fourthly we will provide graphical representations of lifecycles that characterize low and high quality articles from our German Wikipedia test base. On the basis of these lifecycles we will extract our metrics for measuring quality. Fifthly we will evaluate these metrics and compare them to others currently being discussed in the research field. Finally we will summarize our results and provide a preview of our work to come.

2. RELATED WORK

The incredible success of Wikipedia has attracted a lot of researchers. So it is not surprising that numerous publications about Wikipedia have appeared in the last few years. There is a wide and interdisciplinary array of issues being discussed, such as visualization tools [20, 21, 16], motivations for participation [8], the effects of coordination and collaboration [24], vandalism analysis and detection [10, 15, 17, 19], reputation systems [1, 13, 25], quality assurance and automatic quality measurement [1, 3, 4, 6, 12, 13, 18, 25]. Relating to quality assessment there are two divisions of research. The first group investigates the trustworthiness of the text of a Wikipedia article whereas the second one is involved in the assessment of the quality of the article as a whole.

2.1 Computing the Trustworthiness of Text

The methods in this category offer a means for predicting the accuracy of some facts of an article. *Cross* [4] introduces an approach that calculates the trustworthiness throughout the life span of the text in the article and marks this by using different colors. *Adler* and *de Alfaro* calculate the reputation of the authors of the Wikipedia by using the survival time of their edits as the first step [1]. Then they analyze exactly which text of an article was inserted by precisely which author. Finally, based on the reputation score of the respective authors, Adler and de Alfaro are able to compute the trustworthiness of each word [2]. Analog to Cross they illustrate the trustworthiness by using color-coding.

2.2 Assessing the Quality of Articles as a Whole

A first work in this category was published by *Lih* [12], who discovered a correlation of the quality of an article with the number of editors as well as the number of article revisions. *Lim et. al.* [13] define three models for ranking Wikipedia articles according to their quality level. The models are based on the length of the article, the total number of revisions and the reputation of the authors, which is measured by the total number of their previous edits. *Zeng et al.* [25] propose to compute the quality of a particular article version with a *Bayesian network* from the reputation of its author, the number of words the author has changed and the quality score of the previous version. Furthermore, on the basis of a statistical comparison of a sample of Featured and Non-Featured Articles in the English Wikipedia, *Stvilia et. al.* [18] constructed seven complex metrics and used a combination of them for quality measurement. *Dondio et. al. [6]* derived ten metrics from research related to collaboration in order to predict quality. *Blumenstock* [3] investigates over 100 partial simple metrics, for example the number of words, characters, sentences, internal and external links, etc. He evaluates the metrics by using them for classifications between Featured and Non-Featured Articles. *Zeng et. al.*, *Stvilia et. al.* and *Dondio et. al.* used a similar evaluation method which enables the evaluation results to be compared. Blumenstock demonstrates, with an accuracy of classification of *97%*, that the number of words is the best current metric for distinguishing between Featured and Non-Featured Articles.

2.3 Research Questions

The surprisingly high accuracy of the word count implies that there may be no need to investigate in better metrics. We suggest that with improved evaluation methods, significant evidence of the benefits of using studied metrics can be obtained and that the accuracy of the categorization will be reassessed.

It is expected that the use of Non-Featured Articles as examples for low quality articles deters from the evidence of the evaluation. We believe that a particular portion of Non-Featured Articles is of high quality too. However, this category includes a large number of short articles. An exploration of the German Wikipedia from *January 2008* shows that about *50%* of the articles contain less than *500* characters compared to an overall average of about *2,000* characters. So it can be assumed that some short Non-Featured Articles are of high quality since their subject matter can be briefly but precisely explained.

Furthermore, when an article obtains the featured status, it is displayed on the respective pages and thereby attracts many more web users for contribution. We assume that this fact positively influences some metrics studied for quality assessment, in particular the length, concluding that a high word count in these articles is expected. Our investigation of the German Wikipedia reveals this assumption to be true. For example above *95%* of the Featured Articles have an increasing editing intensity after the articles gets the featured status.

To further improve the quality assessment, a weakness in the word count as a quality measure needs to be looked at in terms of robustness. A high quality score can easily be simulated by simply inserting text into an article. Because of the explained

facts, we see growing interest in new, efficient and robust metrics for quality measurement.

3. USER-DRIVEN QUALITY EVALUATIONS IN WIKIPEDIA

To increase the trustworthiness of the quality, Wikipedia introduced the voting-based quality evaluations *"Articles for Deletion*", "*Good Articles*" and "*Featured Articles*". For the rest of the paper, we refer to them as Wikipedia evaluations. As we investigate the German Wikipedia in this study, we describe how the Wikipedia evaluations are used in the German Wikipedia [22]. The evaluation procedures are similar to the English Wikipedia. First, for all of the Wikipedia evaluations, any user can nominate an article by listing it on the respective nomination site (*Articles for Deletion*, *Candidate for Good Article* and *Candidate for Featured Article*). When an article is nominated, the article is flagged with a special tag. According to the type of evaluation, there are particular criteria that are used for the decision. Featured Articles have the highest quality standard. They have to be accurate, complete and well written. Good Articles are also high quality articles, however, slight inconsistencies in the quality are tolerated, such as a lack of illustrations or small weaknesses in the writing style. Articles for Deletion are articles of particularly low quality that have been tagged for deletion. Criteria are, for example, an unsuitable representation or a lack of relevance for an encyclopedia. However, even Articles for Deletion actually maintain a minimum standard of quality. The articles that are generally uncontroversial for deletion, such as those victimized by vandalism or other nonsense, are deleted quickly by using the speedy deletion procedure.

After the nomination of an article, the community decides via a vote as to whether or not the article complies with certain criteria. The voting period and the voting rule depend on the kind of evaluation. For example, in order to become a Featured Article, a voting period of 20 days and a slight modification of the two-third voting rule are necessary. After a successful election, the Featured and Good Articles are marked by special tags and are displayed in the respective sections of the Wikipedia portal, whereas Articles for Deletion are deleted by an administrator.

4. NOTATION AND METRICS FOR LIFECYCLE CALCULATIONS

To calculate the lifecycle of Wikipedia articles we construct two metrics, the persistent contribution and the transient contribution. In this section we present our Wikipedia model for analysis first. Based on this model we describe the meanings and measurements of the persistent and transient contributions.

4.1 Modeling Wikipedia

Wikipedia includes a great number of articles $i=1..n$ that were edited by the Wikipedia authors during the life span. With every contribution, a new article version $v_{i,j}$ is created. The index i refers to the article identification number and the index j to the version. The versions are chronologically ordered, starting with $j=1$. The version $v_{i,0}$ is technically defined as an empty one, in other words it is the version before any content was added. To analyze the changes over time, the life span has been divided into periods. As the periods of analysis, we use months, since a shorter period causes overly high volatility of the metrics, whereas a longer period does not able us to track the metrics precisely. The period in which a version was generated is called $p(v_{i,j})$. If an article i gets a Wikipedia evaluation, we call the period in which the article becomes a candidate for the respective Wikipedia evaluation $c(i)$.

For the calculation of the persistent and the transient contributions we have to parameterize the differences between two article versions. Therefore, we define the editing distance $dis(i,j,k)$ as that which shows the difference between the versions $v_{i,j}$ and $v_{i,k}$. It refers to the number of words which were deleted from the former version and the number of words which were inserted into the newer version. For the computation of the difference between the versions, we use the commonly known algorithm from Hunt and McIlroy [9], which is also used in the UNIX "*diff*" program. The algorithm is based on the longest common subsequence of two documents. According to the algorithm, replacements of text are interpreted as a deletion and insertion of text.

4.2 Persistent and Transient Contribution

The two metrics that we developed, the persistent and the transient contributions, are used to measure the editing intensity. To determine the transient contribution, we aggregate all contributions that were contributed and reverted in the same time period. These contributions have a short life span and do not improve the value of an article. The unaccepted contributions, such as those due to vandalism, edit wars or other obvious inaccuracies are reverted in a short period of time – typically in less than three minutes [20, 21] – are comprised in the transient contributions. On the other side, the persistent contribution refers to contributions that remain in the article beyond the time period. Due to their life time it is assumed that these contributions are reviewed and generally accepted by the Wikipedia community. We believe that low and high quality articles are particularly different according to their persistent and transient contributions, thus we conclude that these measurements are highly relevant for quality assessment.

To compute the persistent contribution we measure the editing distance between the last article version in a given period and the last one in the previous period. The index of the last version of an article i in a period p we call

$$x(i,p) = max\ x \mid p(v_{i,x}) <= p \qquad (1)$$

Accordingly the persistent contribution is defined as

$$C^{per}_{i,p} = dis(i,x(i,p-1),x(i,p)) \qquad (2)$$

For the calculation of the transient contribution of an article i in a period p we first add the editing differences of all versions in a given period to the respective previous version. Next we subtract the persistent contribution $C^{per}_{i,p}$ from this value. In a period without any edits $(x(i,p)=x(i,p-1))$ we define the transient contribution as 0.

$$C^{tran}_{i,p} = \begin{cases} C^{tran}_{i,p} = 0 & |if\ x(i,p) = x(i,p-1) \\ C^{tran}_{i,p} = \sum_{j=x(i,p-1)}^{x(i,p)-1} dis(i,j,j+1) - C^{per}_{i,p} & |else \end{cases} \qquad (3)$$

Figure 1 shows a fictive example of an article and demonstrates our approach. The illustrated article includes four versions. The first one, $v_{i,0}$, is defined as an empty version and belongs to period 0. The other versions refer to period 1. The contribution in version

$v_{i,1}$ as a valuable contribution is represented with the persistent contribution $C^{per}_{i,1}=9$. The article versions $v_{i,2}$ and $v_{i,3}$ are transient contributions. In version $v_{i,2}$ the article is victimized by vandalism and in version $v_{i,3}$ this contribution is reverted. The following calculation of the transient contribution conforms to the number of words changed in the versions $v_{i,2}$ and $v_{i,3}$.

$$C^{trans}_{i,1} = dis(i,0,1) + dis(i,1,2) + dis(i,2,3) - C^{per}_{i,1}$$

$$C^{tran}_{i,1} = 9 + 2 + 2 - 9 = 4$$

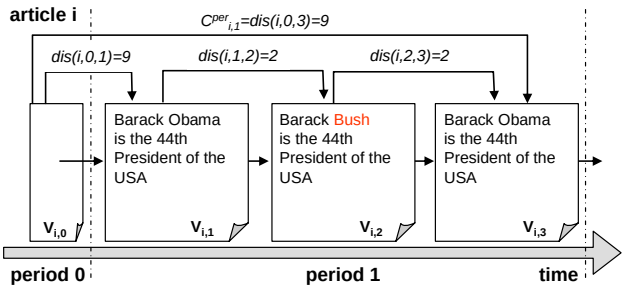

Figure 1. Persistent and transient contribution

Our proposed method to identify unaccepted contributions avoids a common problem. Other published approaches search for edits, which revert a previous version, by parsing for keywords like "*reverted*" or "*vandalism*" within the editing comments or by comparing hash values of article versions [10, 15]. It is assumed that the reverted contributions are unaccepted ones. However these approaches are not able to detect unaccepted contributions completely, if the editors do not comment their edits adequately or if a reversion is combined with other changes. As a weakness of our approach the measures can become tainted if the last versions in the given periods are unrepresentative ones, for example, if they are created by malicious behavior. However, we assume a low probability of such an incident since unrepresentative article versions are removed quickly [20, 21].

Alternatively, the editing intensity in a given period can be measured by simple metrics like the *number of editors* or the *number of revisions* within the period. As we show in section 6 these simple metrics are less effective for quality assessment than the persistent contributions. Hence we do not consider them for the calculation of the lifecycles in the next section.

5. LIFECYCLE ANALYSIS

In this section we describe the dataset we used and we present the methodology that we applied for the computation of the lifecycles. Furthermore we illustrate the extracted lifecycles of low and high quality articles and deduce potential metrics for quality measurement.

5.1 Dataset and Methodology

The database of the common Wikipedias can be downloaded as an SQL-dump from the website. The dump includes the source texts of all pages, consisting of the article text along with the HTML and wiki code and even the complete page edit history. In addition to the text of the article versions, the edit history contains further metadata such as the username of the editors of the revision and the editing time. Wikipedia pages can, however, be anonymously edited without a previous registration or login. For such cases, instead of using the username, the respective IP address is stored. For our investigation we downloaded the database of the German Wikipedia from 01/21/2008 and imported it into a mySQL-Database. We only take into account the pages in the main namespace, which is the namespace for the encyclopedic articles. Pages of other namespaces, for instance user portals, pages of the Wikipedia namespace and discussion pages have been excluded, since we assume that these pages have an uncommon way of editing.

To distinguish between low and high quality articles, we use, as our common procedure, the Wikipedia evaluations described in section 3. As Good and Featured Articles meet similar criteria and as we discover an alternating effect between both types, we consider Good as well as Featured Articles as examples for high quality articles. As discussed in section 2 some other publications assess the quality of Non-Featured Articles as low. In contrast, because the quality of Non-Featured Articles is not clear, we decided to consider Articles for Deletion as examples for low quality articles.

Alternatively, Zeng et. al. utilized *clean-up tags* to identify articles of low quality. These tags point out a variety of areas where quality lacks, such as missing citations and footnotes or an excessive using of jargon or buzzwords. However, we suppose that some cleanup-articles do result in high quality, since the tagging via clean-up tags shows that the Wikipedia community cares about the accuracy and improvement of the respective articles. Hence we disregard clean-up articles in our analysis.

To extract articles in the categories that were explained within the total dataset, we parse the last version of all articles based on their flagging tags. The numbers of articles we found in the different categories is shown in table 1.

Table 1. Article statistics

Quality category	**Low quality**	**High quality**	
Type of evaluation	Article for Deletion	Good Article	Featured Article
Number of articles	147	2184	1211
Sample size / ratio	100 (68%)	50 (2,3%)	50 (4,1%)

The computations of the persistent and transient contributions, in particular the calculation of the editing distance *dis(i,j,k)*, are based on a complex algorithm. Hence the calculations are extremely time-consuming, particularly for articles with a great number of revisions and long text. To handle this difficulty we randomly selected a sample of 100 low quality articles and 100 high quality articles (50 Good and 50 Featured Articles).

As illustrated in Section 2.3 the Wikipedia quality evaluations influence the editing intensity of an article due to the listings on the respective pages. To ensure that our analysis is not affected by this, we truncate the edit histories in the sample after the last article version in the month before the article was nominated for the respective evaluation. Some Featured Articles had the Good Article status before they became Featured Articles. In this case we cut the edit history before the articles became a Candidate for Good Articles. We suppose that the last version of the truncated

edit histories have a comparable quality level as the final version. Thus the calculated lifecycles include the complete development process of low and high quality articles respectively. For the cutting we determined the evaluation status (*Article for Deletion*, *Candidate for Good Article*, *Good Article*, *Candidate for Featured Article* and *Featured Article*) of all revisions in the given sample by parsing the source text for the belonging tags. Afterwards we are able to identify the month before an article was nominated *(c(i)-1)*.

To calculate the lifecycles we compute the metrics for every month in the truncated data for all articles of the sample as explained in section 4.2. For the detection of the characteristics of low quality and high quality articles in regard to their lifecycles we aggregate the measures of both quality categories by averaging. We aggregate the articles according to their maturity, which refers to the amount of time until the nomination month. In other words measurements with the same maturity are summarized in one period of the aggregated life cycle. The number of existent measurements in a period increases with the growing maturity since the articles in the sample have different life spans due to the various creation and nomination dates. Therefore the number of existent articles in a period has to be computed before averaging. In figure 2 our approach for aggregation is illustrated by a fictive example.

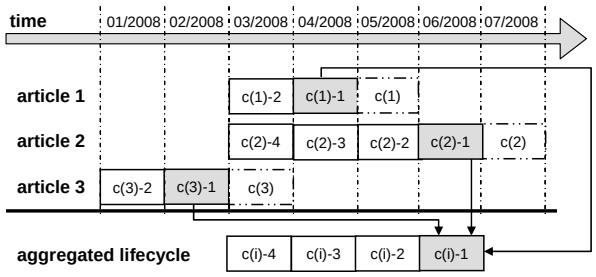

Figure 2. Aggregation of measures

We experimented with two other approaches of aggregation. First, we aggregated the articles according to their age. Thus, in the first period of the aggregated lifecycles the measures of the creation months are averaged. Furthermore we aggregated articles according to the real date. In comparison to the aggregation according to the maturity, the lifecycles calculated in these two ways do not show significant differences between low and high quality articles. Therefore we confine our investigation in the next section to the aggregation according to the maturity

5.2 Lifecycles and Potential Metrics for Quality Measurement

In figure 3 we present the calculated lifecycles of low quality and in figure 4 of high quality articles. The graphs show significant differences between both quality categories. For low quality articles the persistent contribution tends to decrease with increasing maturity. It seems that low quality articles are edited in particular within the early periods whereas only small additions and corrections are accomplished as the article matures. The curve of the transient contribution shows a volatile development probably caused by vandalism. The lifecycle of high quality articles shows a significantly different evolution of the metrics. With the exception of some outliers at the beginning of the lifecycle, the persistent contribution rises with increasing maturity. Particularly in the last three months, the persistent contributions increase greatly. The transient contribution runs in the same way.

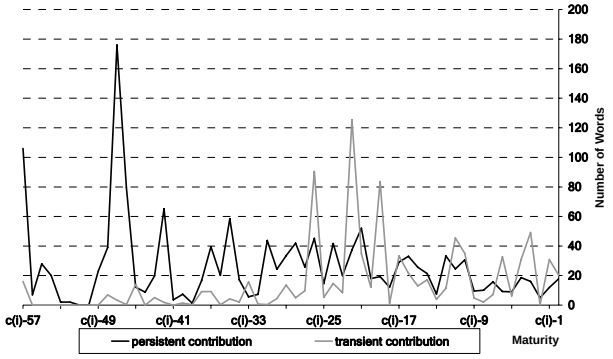

Figure 3. Lifecycle of low quality articles

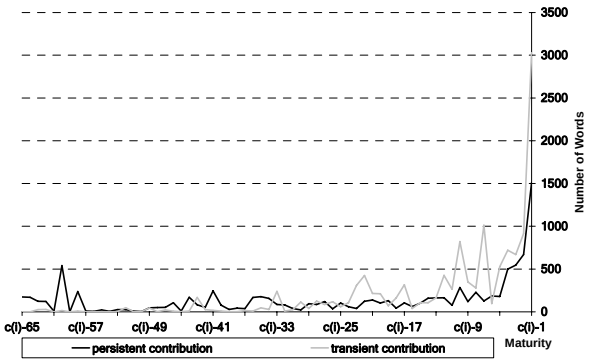

Figure 4. Lifecycle of high quality articles

It appears that at a particular point in time, the high quality articles become the focus of the Wikipedia community and after a stage of intensive editing, the articles become Good or Featured Articles. We cannot explain this extreme increase in the editing intensity with precision. We investigated the article versions in the last three month in detail. We realized that *25* articles of *100* high quality articles were in a review process in these months. The review conduces to quality assurance. Articles in the review are listed at the review page, in which the users can enter their suggestions for improvement. Therefore an increasing editing intensity is expected for these articles. To explore this effect in detail, we computed the lifecycle and left out the articles in the review. The resulting lifecycle was similar to the primary calculated one. It seems that the Wikipedia community is implicitly attracted by listings of the recent changes, watch lists or by talks on the discussion pages. As a result of the different developments of low and high quality articles, particularly in the last three month before the nomination, the maximum values measured throughout the entire life span vary drastically in the quality categories. For example, the maximum measured value for the persistent contribution for the high quality articles is about *1,500* words, whereas, for the low quality articles, it is only about *65*.

As another characteristic of high quality articles, the graph shows that at the end of the lifecycle the transient contribution exceeds

the persistent contribution. The trend suggests that with increasing maturity, the acceptance for new contributions within the Wikipedia community declines, so that at the end of the lifecycle, when the articles are of high quality, a lot of changes are reverted quickly.

To measure the characteristics of low and high quality articles in regard to the different developments of the persistent and transient contribution, we propose in table 2 the following metrics for quality measurement:

Table 2. Potential metrics based on the development of the editing intensity

Metric	Description	Calculation rule
$C^{per}{}_3$	Sum of the persistent contributions in the last three months before nomination	$= \sum_{p=c(i)-3}^{c(i)-1} C^{per}{}_{i,p}$
$C^{tran}{}_3$	Sum of the transient contributions in the last three months before nomination	$= \sum_{p=c(i)-3}^{c(i)-1} C^{tran}{}_{i,p}$
M^{per}	Maximum persistent contribution overall	$= \max_{p=1}^{c(i)-1} (C^{per}{}_{i,p})$
M^{tran}	Maximum transient contribution overall	$= \max_{p=1}^{c(i)-1} (C^{tran}{}_{i,p})$
Q^{per}	Quotient of the average persistent contributions within and before the last three months until nomination	$= \frac{avg_{p=c(i)-3}^{c(i)-1} (C^{per}{}_{i,p})}{avg_{p=1}^{c(i)-4} (C^{per}{}_{i,p})}$
Q^{tran}	Quotient of the average transient contributions within and before the last three months until nomination	$= \frac{avg_{p=c(i)-3}^{c(i)-1} (C^{tran}{}_{i,p})}{avg_{p=1}^{c(i)-4} (C^{tran}{}_{i,p})}$
Q^3	Quotient of the sum of the transient contributions and the sum of the persistent contributions within the last three months until nomination	$= \frac{\sum_{p=c(i)-3}^{c(i)-1} C^{tran}{}_{i,p}}{\sum_{p=c(i)-3}^{c(i)-1} C^{per}{}_{i,p}}$

In addition to the development of the metrics, the graphs reveal differences in the editing intensity in general. The high quality articles show an evidently higher editing intensity throughout the entire life span, measured by the persistent as well as the transient contribution. For example, with the high quality articles, the persistent contribution roughly fluctuates in the majority of the periods between *75* and *200* words per month, with the low quality articles, on the other hand, between *20* and *40* words per month. In the early periods, where the quality of the high quality articles is expected to be low too, the measures are similar.

Due to the differences in the editing intensity in general, additional metrics for quality measurement are defined in table 3.

Table 3. Potential metrics based on the editing intensity in general

Metric	Description	Calculation rule
C^{per}	Sum of the overall persistent contributions	$= \sum_{p=1}^{c(i)-1} C^{per}{}_{i,p}$
C^{tran}	Sum of the overall transient contributions	$= \sum_{p=1}^{c(i)-1} C^{tran}{}_{i,p}$
A^{per}	Average overall persistent contributions	$= avg_{p=1}^{c(i)-1} (C^{per}{}_{i,p})$
A^{tran}	Average overall transient contributions	$= avg_{p=1}^{c(i)-1} (C^{tran}{}_{i,p})$

6. EVALUATION

The lifecycles presented in the previous sections are based on average measures, so that the relevance for the total sample is still unknown. In this section we evaluate the proposed metrics. First, we explain the evaluation method being used in detail, then we offer the evaluation results and discuss the robustness of the metrics, since it is an important feature for practical implementation.

6.1 Evaluation Method

As a commonly used method we judge the metrics by using them for distinguishing between low and high quality articles identified by the Wikipedia evaluations [3, 6, 18, 25]. By comparing our classifications with the given Wikipedia evaluations, we are able to determine the accuracy of the categorizations and hence the significance of the metrics. In contrast to the lifecycle analysis, the metrics are investigated for each article individually.

We evaluated all metrics developed in the previous section. Furthermore we considered the length of an article, *L* as a benchmark, as the most effective metric currently known. *L* refers to the number of characters of the source text of an article version, thus HTML and wiki-code are also included. We measured the length of the last article version in the truncated edit histories. The index of this version is formally defined as:

$$max\ x \mid p(v_{i,x}) < c(i).$$

Moreover, we are interested in the performance of simple metrics to measure the editing intensity in comparison to the persistent and transient contributions. Therefore, in analogy to the developed lifecycle metrics, we also considered the number of editors in the three last periods before the nomination, E^3, the maximum number of editors per period, M^e, the quotient of the average number of editors within and before the last three month before $c(i)$, Q^e, the average number of editors per month, A^e, and the overall number of editors E. However, the number of editors in a given period cannot be determined precisely due to the possibility of anonymous contributions. Hence we predict the measure with the number of distinct usernames and IP addresses in a given period of time.

Within the evaluation we used the same sample of data as that which was utilized for the lifecycle analysis. Due to the small sample size we randomly selected a second sample. According to the similarity of the evaluation results in both samples we are able to verify the significance of the evaluation. The second sample has the same size and ratio of Articles for Deletion (*100*), Good (*50*) and Featured Articles (*50*) as the first sample. For the category of the high quality articles we only considered articles that were not in the first sample. In the total dataset there are only 147 Articles for Deletion, thus duplicates could not be excluded within the category of low quality articles.

For the categorization a simple threshold based classification rule is utilized. We randomly split each sample into two parts, *50%* of the articles in every quality category (*50 Article for Deletion, 25 Good Articles* and *25 Featured Articles*) we used for training, to calculate an appropriate threshold, τ, and the rest for testing.

Table 4 shows some descriptive statistics of the investigated metrics. The median and the standard deviation, σ, are based on the articles of both samples.

Table 4. Descriptive statistics

Metric	Median low quality	Median high quality	σ low quality	σ high quality	τ Sample1	τ Sample2
$C^{per}{}_3$	0	1631	80	2537	209	145
$C^{tran}{}_3$	0	319	333	9873	13	9
M^{per}	151	1947	346	1929	820	662
M^{tran}	28	761	768	9270	103	197
Q^{per}	0	1,60	2,03	379,44	0,93	0,15
Q^{tran}	0	1,28	103,22	92,90	0,04	0,22
Q^3	0	0,21	32,01	6,61	0,01	0,01
C^{per}	258	3735	612	5563	995	1268
C^{tran}	34	1315	1360	16931	109	67
A^{per}	14	181	31	787	64	68
A^{tran}	3	60	64	753	25	27
L	1564	20125	3008	18880	7293	3478
E^3	0	12	2	25	4	4
M^e	4	17	3	14	7	8
Q^e	0	1,59	1,67	3,61	0,77	0,64
E	11	45	17	136	18	40
A^e	0,52	2,19	0,63	2,97	0,80	1,43

The medians of all metrics suggest a positive relation between the quality and the measures. Hence we classified an article as low quality if its measure is less than the given τ; otherwise the article is classified as high quality. The threshold τ was determined via a brute-force search. We reviewed all values for τ between the minimum and maximum measures in the sample. For the continuous measures *(Q^{per}, Q^{tran}, Q^e, Q^3, A^{per}, A^{tran} and A^e)* we tested with a numbers rounded to two decimal places. The threshold τ is defined as the value that achieved the highest accuracy for categorization within the training data. Finally, in order to judge our metrics, we classified the test samples with the calculated τ.

6.2 Results

In table 5 we present the results of our evaluation. The table shows the average accuracy of the categorization as well as the average False Positive rate (FPR - Ratio of the incorrectly categorized low quality articles) and the average True Positive Rate (TPR - Ratio of the correctly discriminated high quality articles) over both samples. Furthermore we present the accuracy in each sample. The comparable accuracies and in particular rankings of the metrics in both samples suggests that our evaluation is significant.

Table 5. Evaluation results

Metric	Rank	Accuracy	TPR	FPR	Accuracy Sample 1	Accuracy Sample 2
M^{per}	**1**	**87%**	76%	3%	84%	89%
A^{per}	**2**	**86%**	78%	6%	85%	87%
M^e	**2**	**86%**	80%	8%	84%	88%
C^{per}	**4**	**85%**	78%	9%	80%	89%
E^3	**5**	**83%**	79%	13%	85%	81%
L	**6**	**82%**	77%	13%	79%	85%
$C^{per}{}_3$	**6**	**82%**	67%	4%	76%	87%
A^e	**6**	**82%**	80%	17%	79%	84%
$C^{tran}{}_3$	**9**	**80%**	74%	15%	78%	81%
Q^3	**10**	**77%**	78%	24%	75%	79%
A^{tran}	**10**	**77%**	67%	13%	74%	80%
C^{tran}	**12**	**72%**	76%	33%	71%	72%
M^{tran}	**13**	**71%**	67%	25%	71%	71%
E	**14**	**70%**	57%	17%	71%	69%
Q^{tran}	**15**	**64%**	45%	17%	70%	58%
Q^{per}	**16**	**63%**	45%	19%	65%	61%
Q^e	**17**	**57%**	61%	37%	66%	58%

The evaluation ascertained high effectiveness for quality measurement of the lifecycle based metrics, in particular of metrics regarding the persistent contribution *(M^{per}, A^{per}, C^{per}, $C^{per}{}_3$)*. With a degree of *87%* for M^{per}, followed by *86%* for A^{per} we could achieve the highest accuracy over both samples. In general, for the metrics based on the persistent contribution, the FPR between *3%* and *9%* is extremely low. The evaluation confirms the assumption that on the one side high quality articles are in general more persistently edited than low quality articles and that on the other side they have a stage of a high editing intensity in their lifecycles. An exception according to the persistent contribution based metric is $C^{per}{}_3$. This measurement only obtains high accuracy of categorization for the second sample.

Furthermore, the analysis verifies that the length of an article is highly relevant for quality measurement. In the second sample L achieves the highest TPR with *90%*. However, our study proves that several lifecycle based metrics are lightly more evident than L. In both samples A^{per}, M^{per}, M^e and C^{per} are slightly more efficient according to the accuracy regarding to L. A detailed analysis of the classification shows that with these metrics some

articles can be judged correct, that are distinguished false with *L* whereas the reverse case appears rarely.

The metrics relating to the last three periods of the lifecycle ($C^{per}{}_3$, $C^{tran}{}_3$, Q^3 and E^3) achieve acceptable accuracy rates. It also confirms that high quality articles normally pass a stage of high editing intensity before they become a candidate for a Wikipedia evaluation. Especially E^3 and $C^{per}{}_3$ with comparable accuracy over both samples like *L* seem to be appropriate for quality measurement. In the first sample M^e attains the highest accuracy in the sample with *85%*. Q^3 with an average accuracy of *77%* over both samples confirms the assumption that for a high portion of the articles the quotient of the accepted (persistent) and unaccepted (transient) contribution reflects the quality. However, as for practical implementation of quality assessment in Wikipedia, articles without a Wikipedia evaluation have to be judged. The period of time in which an article passes the stage of intensive editing before it achieves a high quality level is unknown. In this case the stage of high editing intensity can be identified with M^{per}, for example.

As a further result of our analysis, despite the evident difference between low and high quality articles according to the transient contribution, our analysis reveals that metrics based on the transient contribution *(C^{tran}, $C^{tran}{}_3$, M^{tran} and A^{tran})* are less appropriate for quality assessment. The accuracy over both samples fluctuates, depending on the respective metric, between *71%* and *80%*. In comparison to other metrics, the FPR is particularly poor. For example by using C^{tran} for categorization we measured the second highest FPR in our study with *33%*. As described the transient contribution includes a high portion of contributions like vandalism. It seems that such contributions are more determined by chance than by the quality. The high standard deviation of the transient contribution based metrics (see table 4) confirms this assumption.

Despite the extreme increase of the editing intensity at the end of the lifecycle of the high quality articles, we measured for Q^{per}, Q^{tran} and Q^e the lowest accuracies in our analysis. However, this poor accuracy was strongly caused by the data samples that were used in combination with the evaluation method. To compute these metrics, a minimum life span of four month is necessary. Otherwise the value of the average measure before the last three months is *0*, and a division by zero is produced. In this case we define the metric as *0* too. We discovered some high quality articles in our samples which were nominated shortly after their creation. Accordingly, we truncated the edit histories in an early section. Therefore their life span is too short. If we define in this case the value of the measures for the high quality articles higher than the respective τ, we could increase the accuracy to *79%* for Q^{per}, *82%* for Q^{tran} and *78%* for Q^e. Using this modification, as a benefit, these metrics can correctly judge a high portion of articles that are distinguished false by other metrics. For example about *80%* of the high quality articles, which are judged false with *L*, can be distinguished correctly with Q^{per}, Q^{tran} or Q^e.

Finally, we compared the persistent and transient contribution with simple metrics for the measurement of the editing intensity. The analysis proves that the persistent contribution is in general more effective than the editor count in a period. With the exception of E^3 within the first sample all persistent contribution based metrics exceed their editor count based measures when looking at the accuracy rating. However, the study shows that simple metrics are also appropriate for the measurement of the editing intensity. These metrics achieve in general a higher TPR whereas for the persistent based metrics, a more accurate FPR could be measured. For example with E^3 as well as A^e we could achieve the highest TPR of *80%* in the average of both samples.

6.3 Robustness of the Metrics

Based on the high accuracy of our metrics as compared to other metrics, we suggest the consideration of the persistent contribution based metrics for quality assessment. However, for the practical use, besides the appropriateness of a metric for quality prediction, the robustness of the metrics against manipulations also needs to be considered.

We distinguish between sensitive and insensitive metrics. Sensitive metrics are related to the current version such as *L*. By using these metrics for quality assessment a high quality categorization can be easily inappropriately suggested, for example in the case of *L*, by inserting text. On the other hand malicious contributions such as deletion of text are detected by these metrics.

Our metrics belong to the category of insensitive metrics. They are not directly related to the current version and are based on the edit history. The metrics are not sensitive to changes of the current article. Particularly the maximum persistent contribution can be accomplished in a distant period and the relation to the current version can be lost when an article is no longer in the focus of the Wikipedia community. Otherwise, as an advantage, particularly with the persistent contribution based metrics, they are tamper-proof against quality whitewashing in principle, because they are determined by the acceptance in the Wikipedia community. For example the insertion of haphazard text does not change the persistent contribution, since it is expected that these contributions are reversed rather quickly and are being included in the transient contribution. However, for practical use, our approach has to be modified. We use the last versions of the months to compute the persistent and transient contributions, respectively. By altering an article shortly before the end of the month, our measures can be manipulated. To ensure robustness, we recommend the use of a sliding calculation in relation to the current date, so that the timing of a successful manipulation remains unclear.

As described, both sensitive and insensitive metrics have advantages and disadvantages according to their robustness. As a compromise, in order to match the advantages of both categories, metrics looking at the last three month before the nomination *($C^{per}{}_3$, E^3* and *Q^3)* can be utilized. These metrics are robust against quality whitewashing and can provide information about when an article is no longer gaining so much attention in the Wikipedia community. However contributions like vandalism do not influence the metrics directly. Alternatively, to benefit from the advantages of both sensitive and insensitive metrics, we propose an implementation of quality measurement using various groupings of both types of metrics. In a further piece of work we will investigate the robustness of metrics and the combination of these metrics in more detail.

7. CONCLUSION

This study investigated the automatic quality assessment of Wikipedia articles. In general we could show that Wikis seems to be an appropriate choice for applying automatic quality

measurement techniques. As compared to traditional websites and print media, Wikis offer within the edit history a vast array of information about the development process. This information includes implicit evidence about the quality and thus can be utilized explicitly for quality assessment.

In this paper we first presented the lifecycles of low and high quality articles that we calculated according to the development of the persistent and transient contributions. The study shows significant differences between the two quality categories. The high quality articles are generally more intensively edited and pass through, in contrary to the low quality articles, a stage of extremely high editing intensity before they become a Wikipedia evaluation as either a Good or Featured Article. On the basis of these differences, we constructed 11 metrics for automatic quality measurement. We evaluated these metrics by using them for categorization between low quality *(Articles for Deletion)* and high quality articles *(Good and Featured Articles)*. According to the degree of accuracy we could prove that lifecycle based metrics, in particular those based on the persistent contribution (e.g. the average persistent contribution per month and the maximum persistent contribution throughout the entire life span), are highly effective in quality assessment.

Thus we believe that a practical implementation in Wikipedia of an automatic quality measurement based on these metrics is an interesting and worthy development which can help to judge the trustworthiness of articles. For example the metrics can be used to determine a quality score for each article. According to the very good FPR, there is a high degree of reliability regarding articles classified as high quality. However, our analysis shows a small error rate for all of the investigated metrics. Therefore an implementation of these metrics should include a warning that the quality score can fall short of expectations in certain cases. Alternatively the automatic calculated quality score can be combined with user-driven ratings, as proposed in Kramer et. al. [11].

For further work, first we would like to evaluate other criteria to distinguish between persistent or transient contributions. For example, the number of edits that a contribution survives could be used as an alternative measure instead of the life time. In addition we will investigate in various combinations of the metrics in order to extend our methods for quality measurement. Furthermore we are interested in other reference articles instead of simply the articles judged via Wikipedia evaluations. For example, expert ratings, published in studies that compare Wikipedia with other traditional encyclopedias, could be used. As a potential weakness, also proposed by Blumenstock [3], the validity of Wikipedia evaluations can be discussed. It can be assumed that according to the voting procedure the most popular articles are elected for Good and Featured Articles and maybe not the articles that truly maintain the highest quality standard. Furthermore by using articles assessed by Wikipedia evaluations, the metrics for the time period in which an article is determined to be high quality, cannot be measured in that exact same period of time. The metrics are influenced by the attraction of editors after articles are listed on the respective Wikipedia pages. By using expert rated articles instead, this period of time can be analyzed too, which may provide more efficient metrics. To conclude, we place great value on studying the robustness of our metrics in detail.

8. ACKNOWLEDGMENTS

Special thanks to Lisa Anders for helpful advices and valuable proofreading.

9. REFERENCES

[1] Adler, B.T. and de Alfaro, L. 2007. A Content-Driven Reputation System for the Wikipedia. In Proceedings of the 16th International Conference on the World Wide Web. 261-270, (May, 2007), Banff, Canada.

[2] Adler, B.T., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I. and Raman, V. 2008. Assigning Trust To Wikipedia Content. In Proceedings of the 2008 International Symposium on Wikis. (September, 2008), Porto, Portugal.

[3] Blumenstock, J.E. 2008. Size Matters: Word Count as a Measure of Quality on Wikipedia. In Proceedings of the 17th international conference on World Wide Web. 1095-1096, (April, 2008). Beijing, China.

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

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

[6] Dondio, P. and Barrett, S. 2007. Computational Trust in Web Content Quality: A Comparative Evalutation on the Wikipedia Project. In Informatica – An International Journal of Computing and Informatics, 31/2, 151-160.

[7] Giles, G. 2005. Internet encyclopedias go head to head. In Nature, 438, 7070, 900-901.

[8] Hoisl, B., Aigner, W. and Miksch, S. 2007. Social Rewarding in Wiki Systems – Motivating the Community. In Proceedings of the second Online Communities and Social Computing. 362-371, (July, 2007), Beijing, China.

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

[10] Kittur, A., Suh, B., Pendleton, B.A. and Chi, E.H. 2007. He says, she says: Conflict and coordination in Wikipedia. 25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007). 453-462, (April/May, 2007), San Jose, USA

[11] Kramer, M. Gregorowicz, A. and Iyer, B. 2008. Wiki Trust Metrics based on Phrasal Analysis. In Proceedings of the 2008 International Symposium on Wikis, (September, 2008), Porto, Portugal.

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

[13] Lim, E.P., Vuong, B.Q., Lauw, H.W. and Sun, A. 2006. Measuring Qualities of Articles Contributed by OnlineCommunities. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. 81-87, (December, 2006), Hong Kong.

[14] O'Reilly, T. 2005. What is Web2.0? http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

[15] Priedhorsky, R., Chen, J., Lam, S.K., Panciera, K., Terveen, L. and Riedl, J. 2007. Creating, Destroying, and Restoring Value in Wikipedia. In Proceedings of the 2007 international ACM conference on Supporting group work, 259-268, (November, 2007), Sanibel Island, USA.

[16] Sabel, M. 2007. Structuring wiki revision history. In Proceedings of the 2007 International Symposium on Wikis. 125-130. (October, 2007), Montreal, Canada.

[17] Smets, K., Goethals, B. and Verdonk, B. 2008. Automatic Vandalism Detection in Wikipedia: Towards a Machine Learning Approach. In Proceedings of the AAAI Workshop, Wikipedia and Artificial Intelligence: An Evolving Synergy (WikiAI08). (July, 2008), Chicago, USA

[18] Stvilia, B., Twidale, M.B., Smith, L.C. and Gasser, L. 2005. Assessing information quality of a community-based encyclopedia. In Proceedings of the International Conference on Information Quality, 442–454, (November, 2005), Cambridge, USA

[19] Potthast, M., Stein, B., and Gerling, R. 2008. Automatic Vandalism Detection in Wikipedia. In Proceedings of the Advances in Information Retrieval - 30th European Conference on IR Research. 663-668, (March/April, 2008), Glasgow, UK.

[20] Víegas, F., Wattenberg, M. and Dave, K. 2004. Studying cooperation and conflict between authors with history flow visualizations. In Proceedings of the SIGCHI Conf. on Human Factors in Computing Systems, 575–582, (April, 2004), Vienna, Austria.

[21] Viégas, F., Wattenberg, M., Kriss, J. and Ham, F. 2007. Talk before you type: Coordination in Wikipedia. In Proceedings of the 40th Hawaii International Conference on System Sciences. 78-88, (January, 2007), Hawaii, USA.

[22] Wikipedia. 2009. Autorenportal. http://de.wikipedia.org/wiki/Wikipedia:Autorenportal

[23] Wikipedia. 2009. Multilingual statistics. http://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics

[24] Wilkinson, D. and Huberman, B. 2007. Cooperation and quality in Wikipedia. In Proceedings of the 2007 International Symposium on Wikis. 157-164, (October, 2007), Montreal, Canada.

[25] Zeng, H., Alhoussaini, M., Ding, L., Fikes R., and McGuinness, D. 2006. Computing trust from revision history. In Proceedings of the Intl. Conf. on Privacy, Security and Trust, (October/November, 2006), Markham, Canada.