

The Impact of Heavy Editorial Events on Wikipedia Page Quality

Corey Oliver
The University of Iowa
Iowa City, IA 52242
cjoliver@uiowa.edu

ABSTRACT

Wikipedia is a collaborative encyclopedia effort that allows its users to contribute information on a variety of topics. Considering the size and scope of articles available on Wikipedia, it represents a rich source of research data.

In this paper I study the impact of high editorial events on Wikipedia page quality. First I retrieve a set of pages which have had a high editorial event. A *high editorial event* for some page, is a time interval where the number of revisions contributed is greater than or equal to the number contributions in the previous time interval plus some natural number k . With this set, I determine each page's quality before and after a high editorial event has occurred. To calculate page quality I use methods described in [6]. Finally, I measure the impact different values of k have on page quality before and after a high editorial event.

The data I collected shows that high editorial events contribute positively to a pages quality. Also, high editorial events with higher k values positively impact page quality more so than their lower k counterparts.

1. INTRODUCTION

Wikipedia¹ is a popular collaborative encyclopedia project which enables almost any user of the site to edit and revise its articles. More than ten years after its launch in 2001, Wikipedia has over 23 million articles and more than 14 million registered users [9]. The size and scope of available articles in Wikipedia far outnumber other projects of similar nature, creating a rich environment for webmining studies.

Due to the nature of Wikipedia's collaboration policy combined with the response time of user contribution, Wikipedia allows for articles or article contributions which are temporally sensitive to be included in traditional encyclopedias, an example being current events.

¹<http://www.wikipedia.org>

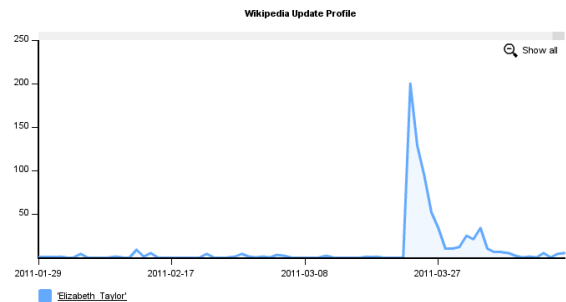


Figure 1: A plot of revision activity of the Elizabeth Taylor Wikipedia page for early 2011. Notice the sharp increase in revisions on March 23, the day of her death. This graph was produced by WikiChanges, which can be accessed at irlab.fe.up.pt/p/wikichanges/.

It is often the case that when news is reported on a specific topic an article may exhibit an unusual increase in its number of contributions. Take for example the death of a celebrity. In the case of the recent passing of Elizabeth Taylor, there were 200 contributions on the day of her death (March 23, 2011). In stark contrast, the most recent contribution to the page occurred almost a week prior on March 18 and was the only instance occurring for that day.

Spurred by the death of Elizabeth Taylor, many users submitted revisions which altered general aspects of the article unrelated to her death. One would expect that if a topic receives above average attention its overall quality would increase.

In this paper I explore the impact high editorial activity has on Wikipedia article quality. The general idea is that a Wikipedia article can be defined to have had *high editorial activity*, if at some time interval in the article's revision history there has been a significant increase in user contributions when compared to the previous time interval. I term the time interval where this increase occurs as a *high editorial event*. Using this definition, the notion of a high editorial event is relative to the amount of contributions that are made to an article in general. In light of this, I determine if a given article has possessed high editorial activity (ie. contains a high editorial event) using the following process:

- I extract the entire revision history of an article and split-up these revisions into partitions based on the time interval the a revision took place. For example, if the time interval has a length of 1 day, then the first revision partition would contain all revisions which occurred between when a page was created and 1 day later. The second revision partition would span the day after this and so on.
- I construct a function which takes in a natural number i and outputs the size of the i -th revision partition.
- I finally find the derivative of this function, which I use to obtain if the article has had a heavy editorial event.

If an article is found to have had a heavy editorial event, I compare the quality of the page before and after the event as described by the authors of [6].

My primary motivation for choosing this topic is to discover whether an high editorial event contributes to a higher quality Wikipedia page. I hypothesize that this is the case.

2. RELATED RESEARCH

The enormous size and scale of Wikipedia has generated significant interest in the research community. In this section I give a synopsis of various works that have contributed to the methods and ideas used in this paper.

The analysis of trends of article revisions has also been a topic of interest for researchers. Several studies have examined article content and how it changes through time. The approach by early researchers in this area was to measure to simply word-based text differences between revisions [1][7]. More sophisticated analyses have distinguished categories of content (such as body text, images, references, etc.) and used this additional information when calculating revision difference [4]. For example, the authors in [4], utilize this additional information to determine if an edit is *significant*. An example of significant edit, according to [4] would be a user modifying the structure of an article or adding new content. A less significant edit would be the a simple spelling or grammatical fix.

The area of automatic quality assessment for Wikipedia pages has also been intensely studied. One of the first works to study this in the more general environment of the web was [8]. In this paper the author attempts to devise a scheme to determine quality of arbitrary web pages. His work has been used as a basis for similar research in the Wikipedia domain [2][3]. However, the nature of Wikipedia’s open contribution scheme exhibits some stark differences from typical web pages. Namely, Wikipedia lacks stability and control over content quality. Some authors have proposed solving this problem by treating it as classification problem [5]. Others have studied the problem using different methods. In this paper I use definition of quality developed by Javanmardi et al[6]. They utilize a combination of user authority, the quality of past user contributions, and other metrics to define an article’s quality. This is discussed in greater detail in section 3.2.

3. METHODS

An issue one must grapple with for any data source of Wikipedia’s size, is how to aggregate data in an efficient manner. Wikipedia facilitates this with their API², which allows users to extract information about numerous aspects of Wikipedia, including revision data, user attributes, and article content. Along with the Wikipedia API, I use another API³ provided by the authors of [6]. This is used to gather data needed to calculate page qualities.

3.1 Choosing Articles to Study

Let W be the set of all articles on Wikipedia. I first collect a set of Wikipedia articles which have exhibited above normal editorial activities for some time interval. Let’s call this collection HW . I determine this collection using the method described below.

I begin by enumerating all revisions R_w for a given article $w \in W$ and partitioning these revisions based on a set of non-overlapping time intervals I . I order this set by defining the t -th interval I^t as

$$I^t = \langle i_t, i_t + r \rangle \quad (1)$$

where r is some time range and i_t is the start time of the t -th time interval. The set RS then contains the set of revision partitions where each partition contains all revisions which took place during the time interval I^t . Formally this is defined as follows:

$$RS_w^I = \{R_w^{I^t} | I^t \in I\} \quad (2)$$

where $R_w^{I^t}$ contains all revisions for article w in the time interval I^t . RS_w^I is ordered by defining the i -th revision partition as:

$$RS_{(w,i)}^I = \{R_w^{I^t} | t = i\} \quad (3)$$

Intuitively, this equation is saying that the elements of RS_w^I are ordered by the time interval their revisions belong to. As an example, if one element’s revisions belong to I^1 it would come before the element whose revisions belong to I^2 .

I now want to retrieve the number of revisions in each RS_w^I partition. This is accomplished through the function `numrevs`. Given a positive $i < |RS_w^I|$, `numrevs` is computed as follows:

$$\text{numrevs}(i) = |R_{(w,i)}^{I^t}| \quad (4)$$

I also define the derivative of this function as `numrevs'`.

²<http://en.wikipedia.org/w/api.php>

³<http://nile.ics.uci.edu/events-dataset-api/api.jsp>

Using this, some article w has a heavy editorial event iff there exists some m in the set of the range of numrevs' which is greater than some natural number k . Here k is some user defined value which gives the number of revisions a page must surpass in order to be considered to have heavy editorial activity.

This method described above generates the set HW . For my research I choose 100 articles from Wikipedia which satisfy HW membership. I then study the quality of these pages before and after a heavy editorial event.

3.2 Analyzing Article Quality

I use the methods described in [6] to access article quality in this paper. A brief overview of this process follows.

According to [6], a Wikipedia article at some point in time can be seen as having a *low quality* ($q = 0$) or *high quality* ($q = 1$). Two metrics are used to determine this q value at a given revision: author reputation and whether this revision has been reverted in a subsequent revision.

To calculate a user's reputation, we assume user i has contributed $N_i(r)$ tokens to Wikipedia before revision r . $n_i(r)$ of these tokens are preserved (they have not been deleted or replaced by other users). The user inserts $c_i(r)$ new tokens at revision r in a Wikipedia article w , $g_i(r)$ of these tokens remain in w . The reputation for user i is defined as follows:

$$UR_i^+(r) = \frac{n_i(r) + g_i(r) - \sum_{d=1}^{p_i(r)} UR_{j(r_d)}(r_d)e^{-\alpha(r_d-r)}}{N_i(r) + c_i(r)} \quad (5)$$

where r_d represents the revision at which the corresponding token was deleted. $UR_j(r_d)$ is the reputation of this deleter, and $p_i(r)$ is the number of deleted tokens.

The authors trained $UR_i^+(t)$ over varying values of α with the goal of maximizing the area under the ROC curve. Results from this showed that $\alpha = 0.08$ returned the optimal result.

Next the authors define the function $Q(R)$, which represents the ratio, as of revision R , of high quality revisions contributed to the Wikipedia article:

$$Q(R) = \sum_{i=1}^n q(r_i)/n \quad (6)$$

Here n is number of revision from r_1 to R .

Finally to study the proportion of time during which the article is in a high-quality state, the authors define $QD(R)$:

$$QD(R) = \frac{\sum_{i=1}^n (r_{i+1} - r_i)q(r_i)}{R - t_i} \quad (7)$$

With this foundation we are now able to study the quality

of arbitrary Wikipedia articles. In particular, I will discuss evaluation of the quality of articles in the HW set. To reiterate, the articles present in HW are those which have exhibited heavy traffic at some point in their existence as formally defined in section 3.1.

4. WEBSITE

<http://www.coreyoliver.org/misc/22c196/project.html>

5. RESULTS

In this section, I use my methods to analyze case studies. In particular, I conducted experiments by using English Wikipedia revisions made before January 30, 2010⁴.

The dataset provided by [6] is provided in flat file database ordered by page id. In my experiments I choose the first 100 pages which exhibit high editorial activity while iterating through this dataset. This is why many pages appearing in my results have titles which appear early in the alphabet. Sometimes a page may have multiple high editorial events. If this is the case, I consider the first of these events in my study.

Considering the method for detecting high editorial events shown in this paper, I construct several studies: I analyze the method using different values of k . With respect to k , a page has had high editorial activity if considering two consecutive time intervals I^1 and I^2 the difference between the number of revisions contributed in I^2 and the amount contributed in I^1 is greater than or equal to k . If k is large, this means there has been a sudden increase of revision contributions from one time interval to the next.

In all these studies, I look at how quality changes on a page when a high editorial event occurs. That is, I compare the quality of a page after the time interval of a high editorial event to the quality of the page in the previous time interval. My studies will show whether there are any significant impacts to page quality due to high editorial events.

5.1 Case Studies

As described previously, I study the impact of high editorial activity on page quality by analyzing sets of 100 Wikipedia pages which have exhibited high editorial activity for different values of k . I choose to study time intervals which are 24 hours. This means that the first time interval I look at is January 1, 2001 12:01am to January 2, 2001 12:00am. A starting date of January 1, 2001 is chosen because this is the month articles first began appearing on Wikipedia.

In Table 1 I show the average difference between the quality of a page at the time of a high editorial event and the quality of the page during the previous time interval using different value for k . I also look at the articles which showed the most positive and negative effects after a high editorial event. These can be seen in Table 2 and 3.

The question I asked at the beginning of this paper was whether a high impact event positively impacted the quality of a Wikipedia article. As seen in Table 1, there has been a

⁴This is due to the dataset provided by [6] only containing revision information before this date.

significant increase on average of page qualities. The effect of high-impact events appears less pronounced when lower k values are used. Overall, the data shown in Table 1 supports the hypothesis I put forward in Section 1. This hypothesis predicted high editorial events would contribute positively to a pages quality.

I also looked at outliers in my results by studying pages most negatively and positively impacted by high editorial events. In Table 2, all entries are for pages with high editorial events of $k = 100$. This is expected, as the quality difference for pages with high editorial events of $k = 100$ was significantly higher than its lower counterparts. Table 3 displays a mix of quality differences for pages containing high editorial events of $k = 50$ and $k = 25$. Something interesting to note in this table is the lack of any negative values in the right column. From this, we can conclude no pages studied were negatively impacted by a high editorial event.

Table 1: Using different values for k , the average difference between the quality of a page at the time of a high editorial event and the quality of the page during the previous time interval.

k	Quality Difference
25	0.1533
50	0.1656
100	0.2803

Table 2: The five articles who were most positively impacted by a high editorial event. That is, the pages tested which had the highest quality difference before and after a high editorial event.

Article Title	k	Quality Difference
Art	100	0.3862
Allen Ginsberg	100	0.3821
Anarcho-capitalism	100	0.3791
Chiropractic	100	0.3710
Death	100	0.3679

Table 3: The five articles who were most negatively impacted by a high editorial event. That is, the pages tested which had the lowest (or even a negative) quality difference before and after a high editorial event.

Article Title	k	Quality Difference
Alphabet	25	0.0083
Abraham Lincoln	50	0.0437
Azerbaijan	25	0.0438
Alaska	25	0.0511
Apollo	50	0.0583

6. LESSONS LEARNED

I ran into numerous issues while conducting research for this project. The most difficult of these was dealing with a particular limitation of the Wikipedia API.

For the data in my project, I used two sources: the Wikipedia API and an API provided by the authors of the paper [6]. One part of determining a page’s quality is calculating the

user reputation of each user who contributed a revision to the page. To calculate a user’s reputation, you must study all the contributions they’ve made to Wikipedia.

I encountered a problem when needing to use the Wikipedia API to gather this contribution information. I start by querying the Wikipedia API for user contribution information. Next I query the author’s API for additional information to calculate user reputation. To calculate user u ’s reputation using the [6]’s API, I need to find the reputation of all users which modified u ’s Wikipedia submission. Again, to do this I must query the Wikipedia API for contribution information about that user. However, Wikipedia only allows queries with usernames and the [6]’s API only provides user ids.

I contacted the author’s of [6] concerning my problem and they kindly provided me with an SQL dump which mapped user ids to usernames.

7. CONCLUSION

In this paper I presented an approach for determining whether a given Wikipedia page had a high editorial activity. I leveraged this notion to find whether high editorial events impacted a page’s quality. To calculate a pages quality I used methods described by Javanmardi et. al in [6]. Using a set of pages with high editorial activity and various values of k , I studied how page quality changes before and after a high editorial event. I also provide the Wikipedia pages which exhibited the most negative and positive page quality differences as a result of a high editorial event. In conclusion, I found that on average a high editorial activity significantly increased the quality of a page. Pages with high editorial events of larger k values appeared to have larger quality increases compared to their lower counterparts.

8. REFERENCES

- [1] B. T. Adler, L. de Alfaro, I. Pye, and V. Raman. Measuring author contributions to the wikipedia. In *Proceedings of the 4th International Symposium on Wikis*, WikiSym ’08, pages 15:1–15:10, New York, NY, USA, 2008. ACM.
- [2] P. Dondio, S. Barrett, S. Weber, and J. Seigneur. Extracting trust from domain analysis: A case study on the wikipedia project. *autonomic and trusting*. In *Computing, Proceedings Lecture Notes in Computer Science 4158*, pages 362–373, 2006.
- [3] S. B. P. Dondio and S. Weber. Calculating the trustworthiness of a wikipedia article using dante methodology. IADIS e Society Conference, 2006.
- [4] P. K.-F. Fong and R. P. Biuk-Aghai. What did they do? deriving high-level edit histories in wikis. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, WikiSym ’10, pages 2:1–2:10, New York, NY, USA, 2010. ACM.
- [5] D. Hasan Dalip, M. André Gonçalves, M. Cristo, and P. Calado. Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, JCDL ’09, pages 295–304, New York, NY, USA, 2009. ACM.

- [6] S. Javanmardi and C. Lopes. Statistical Measure of Quality in Wikipedia. In *1st Workshop on Social Media Analytics (SOMA '10)*, July 2010.
- [7] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and M. T. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *Alt.CHI*, 2007, 2007.
- [8] K. H. Veltman. Access, claims and the quality on the internet – future challenges. *Progress in informatics : PI*, 2:17–40, March 2005.
- [9] Wikipedia. Statistics — Wikipedia, the free encyclopedia, 2010. [Online; accessed 28-March-2011].

APPENDIX

A. EXAMPLE OF A HIGH EDITORIAL EVENT

Consider the Wikipedia page for Elizabeth Taylor, and let r , the length of the time intervals we will be using, be 24 hours. Also let i_1 , the start time of the 1st time interval be March 20, 2011. This date is chosen because it is near the time of Elizabeth Taylor’s announced passing.

Now we can create the set RS_w^I , which contains sets of revisions which take place over our chosen time interval. I identify revisions with their revision id. Note that I consider the i -th element of this set to be the $R_W^i \in RS_w^I$ such that t equals i . That is, the elements RS_w^I are ordered by the time interval their revisions belong to. For example, if one element’s revisions belong to I^1 it would come before the element whose revisions belong to I^2 .

$$RS_w^I = \{\{\}, \{\}, \{\}, \{420313605, 420313807, \dots, 420408700\}, \dots\} \quad (8)$$

Notice that there are no revisions for the first three subsets of RS_w^I and the 4th contains several revisions (many of which are not listed). We want to somehow detect this 4th element as a high editorial event. This is done by transforming the RS_w^I into a function and then finding the derivative of this function.

We call the function we create `numrevs`. Given some i the function produces the number of revisions for the i -th element of RS_w^I . For example, `numrevs(4)` when using the definition of RS_w^I above returns 200, `numrevs(3)` returns 0, and so forth.

We now find the derivative of `numrevs` and call it `numrevs'`. Now, when we look for i values such that `numrevs'(i) >= k` for some k . This means we’re looking for time intervals which have k or more revisions than the previous time interval. For our purposes let’s set $k=100$.

Consider `numrevs'(4)`, which equals 200 and is greater than k . This means the difference between `numrevs(4)` and `numrevs(3)` is greater than k . As a result, the set of revisions contained in the i -th element of RS_w^I is a high editorial event. We now determine page quality before and after this event using methods described in this paper.