Analysis of Revision Activities and Patterns of Wikipedia Articles

Abstract

As a successful crowdsourcing project, Wikipedia has been researched extensively since the last decade. To better understand the dynamics of revision activities in the lifecycle of Wikipedia articles, we identify 14 revision actions, annotate 6,950 revisions from 20 articles randomly selected under four quality ranks (C, B, GA, and FA), and analyze flows of revisions and revision actions in 10 time periods from an article's inception to the promotion to its highest quality rank. We find that higher quality (GA and FA) articles demonstrate a significant raising pattern in amount of revisions and some revision actions at the last time period prior to the promotions. On the basis of our derived revision session concept, using association rule mining we identify sets of revision actions that contributors tend to perform together in the same revision sessions.

Keywords: *Crowdsourcing, Collaborative Processes, Wikipedia, Revision Behaviors*

1. Introduction

Crowdsourcing is an effective way of obtaining needed information or services from large population of an online community. The success of crowdsourcing depends on collaborative and knowledge-intensive processes. Thus, crowdsourcing has become a new area to receive academic attentions (Zhao & Zhu 2014). As the largest and most popular wiki-based peer-production knowledge repository, Wikipedia has been widely studied since the last decade. Wikipedia article revision history serves as a valuable resource in order to understand the relationship between revisions and information quality (e.g., Kane 2011; Liu & Ram 2011).

In this article, we identify 14 types of actions for Wikipedia revision activities. When examining the revision history of a Wikipedia article, we annotate revision activities using those 14 action types. It is worth noting that a revision can consist of multiple, different revision actions. We manually labeled 6,950 revisions from a total of 20 Wikipedia articles randomly selected from four different quality ranks, which are *C*, *B*, *GA* (*Good Article*), and *FA* (*Featured Article*) in the order of low to high. We analyze the dynamics of those revision actions along time dimension from the article's inception to the promotion to its highest quality rank. The contributions of this article are in twofold: (1) higher quality (GA and FA) articles demonstrate a pattern of sharp raising in numbers of revisions and some revision actions at the last time period prior to their promotions; (2) on the basis of our derived concept, revision session, we identify a group of frequently co-occurred revision actions. These findings can be operationalized into guidelines for design and moderation of crowdsourcing communities.

2. Literature Review

This section reviews relevant work with respect to Wikipedia article revisions and article quality. Jones (2008) analyzes the user-composed brief revision summaries to label revisions using 11 revision actions for 10 Wikipedia articles - 5 FA-rank and 5 non-FA-rank articles. His study focuses on the collaborative process of knowledge contributions to Wikipedia articles from the linguistic perspective. Daxenberger & Gurevych (2013) identify 21 revision actions and manually annotate 891 revisions from 20 Wikipedia articles - 10 FA-rank and 10 non-FA-rank articles. They compare the actions between FA and non-FA articles and between pre-promotion and post-promotion periods. Some studies resort to automatic approach in revision annotation. For example, with a total of 1,600 articles from the ranks of C, B, GA, and FA, Liu and Ram (2011) study the relationship between collaboration and article quality. By clustering contributors by their revision actions, they discover various collaboration patterns. Li et al. (2014) conduct a lifecycle analysis of the revision behavior on just the FA-rank articles with a corpus consisting of 86% of all 3,819 FA-rank articles (as of March 2013), and report revision patterns at three phases of the

article lifecycle - pre-nomination, pre-promotion, and post-promotion. Different from their work, we examine the revision activities from inception to promotion at finer granularity (i.e., at 10 time points) with a more meaningful activity categorization (14 action types versus 3) on four different ranks of articles, C, B, GA, and FA. Another novelty distinguishing this study from prior studies is that we investigate the actions within revision sessions, which yields insightful information regarding co-occurred revision actions of the same contributors. We design our research following the three dimensions (control, organization, and data flow) in common process mining and analytics studies. This article mainly focuses on the control flow (revision actions).

3. Data Preparations

3.1 Article Selection

A *WikiProject* is a group of contributors who focus on a specific topic area, location, or set of tasks, and work collectively to improve Wikipedia[1]. According to the Wikipedia Assessment project[2], Wikipedia articles have been ranked by quality. Typical ranks (from low to high) include *Stub*, *Start*, *C-class*, *B-class*, *GA-class*, and *FA-class*. In this study, we select articles from four of the ranks (C, B, GA, and FA) to examine their revision actions. The same selection on article rankings can be found in Liu and Ram (2011). Due to space limit, detailed ranking criteria can be found at the assessment page of Wikipedia[2]. For GA and FA assessment, an article must first be nominated as a GA or FA candidate and then be reviewed externally by editors not limited to this article's WikiProject. If the candidate article passes the review, it receives the quality ranking. B- and C-rank articles are granted by editors of the same project without a nomination process (Liu and Ram 2011).

As of September 2015, the English Wikipedia has nearly five million articles[3] and over 2,000 WikiProjects[1]. For the purpose of article selection, we wanted to choose a WikiProject containing a sufficient number of articles under each of the four quality ranks because such selection would reduce the impacts of different projects toward the article quality. After examining several dozens of WikiProjects, we chose the WikiProject *Dogs* and randomly selected five articles from each of the four quality ranks, which made a total of 20 articles. On July 9, 2015 we programmatically collected all revision history for each of the 20 articles from the time of an article's creation to the date when the article received its highest rank.

3.2 Action Identification

Based on prior literature (Liu & Ram 2011; Daxenberger & Gurevych 2013) and our observation on Wikipedia revisions, we identified 14 actions. Table 1 shows these actions and their definitions. For actions 7 to 9 (*A7 – A9*), a Wikipedia internal link or wikilink refers to a link to another Wikipedia article; while an external link refers to a link to a website outside of Wikipedia. After reading prior literature and examining Wikipedia revisions, the authors agreed upon a set of rules to annotate a revision using the 14 revision actions. When labeling actions for each revision, we also counted the frequency of each action, however, if the action *A14* (vandalism) or *A13* (revert) occurs in a revision, we set the frequency as 1. One of the authors manually examined and annotated all revisions for the twenty articles.

No.	Action Name	Action Definition
A1	Sentence Insertion	insertion of a new sentence
A2	Sentence Modification in a paragraph	modification of one or more sentences in a paragraph
A3	Sentence Deletion	deletion of a sentence

[1] https://en.wikipedia.org/wiki/Wikipedia:WikiProject
[2] https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment
[3] https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

A4	Media Insertion	insertion of a new file or image
A5	Media Modification	modification of a file or image or the caption of the file or image
A6	Media Deletion	deletion of a file or image
A7	Link Insertion	insertion of either a Wikipedia internal or external link
A8	Link Modification	modification of either an internal or external link or hypertext of the link
A9	Link Deletion	deletion of either an internal or external link
A10	Reference Insertion	insertion of a new reference
A11	Reference Modification	modification of a reference, e.g., changes in authors, publishers, page numbers, etc.
A12	Reference Deletion	deletion of a reference
A13	Revert	undo and change back the revision to an early version
A14	Vandalism	any addition, removal, or change of content to deliberately damage Wikipedia

Table 1. Our identified revision actions and definitions

4. Analysis and Results

4.1 Descriptive Statistics

Article length and number of revisions

First, we explore the revision data set and briefly report some descriptive statistics. The 20 articles contain a total of 6,950 revisions (empty revisions were excluded), and among the 20 articles the numbers of revisions range from 22 to 1,053. Table 2 illustrates the average article lengths (in bytes) as well as average numbers of revisions by quality ranks. It is intuitive to observe that the article length increases as the quality rank increases. The average numbers of revisions follow a similar pattern, though the B-class is slightly higher than the GA-class. This is because there is an article, *Border Collie*, in the B-class, which contains the largest number (1,053) of revisions in the corpus.

Rank	**Average article length (bytes)**	**Average number of revisions per article**
C	11,932	231
B	28,698	372
GA	34,703	348
FA	38,603	439

Table 2. Average article length and average number of revisions

To monitor the growth of revisions across different time points, we evenly divided the time frame from an article's inception to its promotion into ten periods for each article. t_1, t_2, …, t_{10}, denote the end times of the ten periods. We analyzed our data set from different perspectives on the basis of these time points.

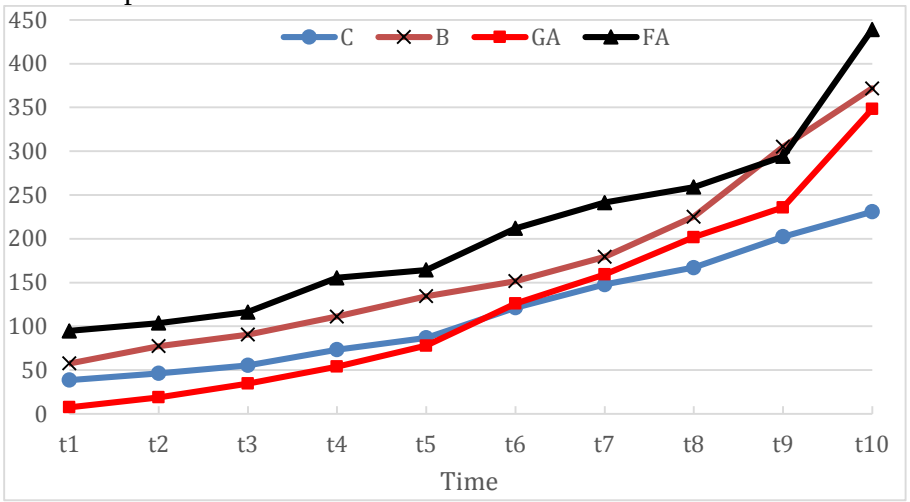

Figure 1. Average numbers of revisions at different time points

Figure 1 illustrates average numbers of revisions per article for each rank at different time points. Compared with articles of the other three ranks, FA articles have the largest numbers of revisions in 9 out of 10 time points. As the amount of the revisions is positively correlated with the article length, we also examined the changes and growth rates in terms of the average number of revisions across time. During the last time period from t_9 to t_{10}, the changes are from 202 to 231 (a 14% increase) for C-class, from 305 to 372 (a 22% increase) for B-class, from 236 to 348 (a 47% increase) for GA-class, and from 294 to 439 (a 49% increase) for FA-class. Close to 50% increases at the last time period for both GA and FA groups happened during article review period and right before a ranking was granted. Actually, we observed the similar patterns in several analyses in our data set.

Time-wise growth rates for numbers of revision actions

Next, we examine this pattern by drilling down from the level of revision to the level of individual revision action type. Table 3 shows the growth rates during the last time period for every action type and the average of all types in the same quality rank. GA and FA articles have much higher growth rates than the B and C articles in 10 out of 14 action types as well as the averages (highlighted in Table 3 for better readability).

Rank	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**	**A10**	**A11**	**A12**	**A13**	**A14**	**Avg.**
C	15%	16%	15%	7%	0%	29%	9%	8%	16%	42%	12%	7%	18%	18%	14%
B	15%	15%	5%	9%	11%	8%	10%	11%	4%	4%	0%	2%	39%	57%	12%
GA	**136%**	**85%**	**83%**	**59%**	29%	27%	**142%**	**101%**	**115%**	**410%**	**590%**	**243%**	12%	7%	**110%**
FA	**39%**	**57%**	**52%**	**105%**	11%	36%	**55%**	**38%**	**66%**	**54%**	**111%**	**66%**	34%	50%	**55%**

Table 3. Revision action growth rates from t_9 to t_{10}

Such evident raisings in revisions and in revision actions clearly suggest that GA and FA articles received significantly increased contributions from contributors prior to their promotions. From the revision action perspective, sentence level actions (*A1 – A3*) represent improvement in text and they are relevant to the GA and FA assessment criteria of "well-written", "comprehensive", and "broad in coverage". Media insertion (*A4*) and link and reference level actions (*A7 – A12*) indicate enrichment of text and they are associated with the criteria of "well-researched" and "verifiable". It is justifiable that our findings align with the ranking criteria of Wikipedia articles.

4.2 Frequent Action Sets

Association rule mining was originally proposed to discover associations of two itemsets in a large transactional database (Agrawal et al. 1993). Given X and Y are two disjoint itemsets, assuming there is an association between X and Y, denoted as X → Y, which means if a transaction contains itemset X, at certain probability it contains Y as well. Two common metrics to measure the strength of an association are *support* and *confidence*.

To apply association rule mining to our data set, we need to first define *revision session*, which is equivalent to the traditional transaction. In a Web context, a *session* refers to the interactions between a user and a website during a specified period of time. The user can be identified by a unique userid or a unique IP address. 30 minutes is a well-adopted timeframe for session timeout - if the user returns to the same website within 30 minutes since his last interaction with the website, his prior session resumes; otherwise, this visit starts a new session for the user.

We resorted to userids and IP addresses in Wikipedia revision history to identify users, and used 30 minutes as our session timeout. Our data set produced 4,272 revision sessions from the 6,950 revisions. The 14 revision actions are equivalent to items in association rule mining for transaction data. We used Weka (Witten and Frank 2005) v3.6 and ran the classic Apriori algorithm on revision session data. With the minimum confidence at 0.5, we adjusted the minimum support

value such that the algorithm produced 10 association rules, and the minimum support value was 0.06. Due to space limit, only top five association rules (with highest confidence) are shown in Table 4, where numbers in parentheses are counts of revision sessions containing the action sets.

No	Association rule	Conf.
1	Reference Insertion A10 (360) ==> Link Insertion A7 (284)	0.79
2	Reference Insertion A10 (360) ==> Sentence Insertion A1 (277)	0.77
3	Reference InsertionA10 (360) ==> Sentence Modification(s) in a paragraph A2 (273)	0.76
4	Sentence Insertion A1, Sentence Modification(s) in a paragraph A2 (513) ==> Link Insertion A7 (342)	0.67
5	Sentence Modification(s) in a paragraph A2, Link Insertion A7 (568) ==> Sentence Insertion A1 (342)	0.60

Table 4. Top five association rules

Interpretation of the association rules can be beneficial for the design and moderation of crowdsourcing online communities. For instance, for the first association rule, when a reference is inserted to the article, at 79% likelihood a link will also be added. This rule could help Wikipedia to improve its user interface design. Currently, under the article edit mode, the icon of (adding) link is located on the top menu bar which is above the article text, while the icon of (adding) reference is beneath the text, making the two icons far apart. To improve user experience, the link and reference icons can be placed closer, so that when a user clicks the reference icon and enters a reference, he can easily locate the link icon and enter the link.

5. Conclusions and Future Direction

On the basis of the 14 revision action types, we first analyzed the dynamics of revision activities in the lifecycle of Wikipedia articles. Across the time periods from the inception to the promotion of an article, we found that higher quality (GA and FA) articles received significantly increased numbers of revisions and revision actions at the last (10^{th}) time period. In addition, we identify frequently co-occurred revision actions which contributors performed in the same revision sessions. This analysis approach could help the Wikipedia and other online communities to improve the ease of use by redesigning the user interface. Future work includes adding the user factor (such as experiences and roles) into analysis. Besides, as Wikipedia talk page allows contributors to communicate for edit improvements and resolve conflicts, the talk page history log can be useful to understand the revision process.

References

Agrawal, R., Imieliński, T., & Swami, A. 1993. Mining Association Rules between Sets of Items in Large Databases. *In Proc. of the 1993 ACM SIGMOD*, Washington DC, USA, pp. 207–216.

Daxenberger, J. & Gurevych, I. 2012. A Corpus-Based Study of Edit Categories in Featured and Non-Featured Wikipedia Articles. *In Proc. of COLING 2012: Technical Papers*, pp. 712–726.

Jones, J. 2008. Patterns of Revision in Online Writing: A Study of Wikipedia's Featured Articles. *Written Communication*, 25(2), pp. 262–289.

Kane, G. C. 2011. A Multimethod Study of Information Quality in Wiki Collaboration. *ACM Trans. on Management Information Systems*, 2(1), article 4.

Li, X., Luo, Z., Pang, K., & Wang, T. 2013. A Lifecycle Analysis of the Revision Behavior of Featured Articles on Wikipedia. *In Proc. of 2013 International Conference on Information Science and Cloud Computing Companion*, Guangzhou, China.

Liu, J., & Ram, S. 2011. Who Does What: Collaboration Patterns in the Wikipedia and Their Impact on Article Quality. *ACM Trans. on Management Information Systems*, 2(2), article 11.

Witten, I.H., Frank, E. 2005. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed., Morgan Kaufmann, San Francisco, CA.

Zhao, Y. & Zhu, Q., 2014. Evaluation on Crowdsourcing Research: Current Status and Future Direction. *Information Systems Frontiers*, 16(3), pp. 417–434.