THE LIFE CYCLE OF CORPORATE WIKIS: AN ANALYSIS OF ACTIVITY PATTERNS

Ofer Arazy*, Arie Croitoru*, Soobaek Jang**
* The University of Alberta
** IBM Corporation
ofer.arazy@ualberta.ca, croitoru@ualberta.ca, sjang@us.ibm.com

Abstract

Following the success of wikis on the internet (e.g. Wikipedia), corporations have begun adopting wikis. Preliminary evidence suggests that wiki is a sustainable collaboration tool and that wikis deployment is experiencing massive success. The objective of this paper is to provide a large scale evaluation of corporate wikis life cycles. We analyze and categorize the temporal activity patterns of more than thirteen thousand wikis in one multinational organization over a 29 months period. This clustering problem poses some unique challenges, and required the development of novel extensions to existing algorithms. We identified four clusters and their prototypical activity patterns. Our findings show that, contrary to what has been suggested in previous studies, most corporate wikis become inactive after a relatively short period, and less than 20% of wikis show continuous activity. Implications for research and practice are discussed.

Keywords: Wiki, corporate, life cycle, activity patterns, clustering,

1. Introduction

Wiki, derived from the Hawaiian-language word for fast, is a web-based content authoring application that is based on the principles of openness, transparency, and peer-based governance (Wagner 2004). Users can jointly edit a wiki page such that at any point in time the most recent page version reflects the cumulative contributions of all users that have edited the page until then. A wiki application can contain many wiki pages. Wikis have already had a profound impact on the Internet, with Wikipedia being the prominent example. An analysis of design principles and primary features of wiki suggest that wikis could be applied to corporate knowledge management and alleviate the bottlenecks associated with knowledge acquisition processes (Wagner 2004; 2006). To date, relatively little is known about the use of wikis within corporate settings. The distinctive affordances of wiki technology give rise to new collaboration forms, suggesting that there is a need to develop new theoretical frameworks to explain wiki-enabled collaboration (Majchrzak 2009). Preliminary evidence suggests that wikis are sustainable (Majchrzak, et al. 2006) and are being adopted at explosive rates (Arazy et al. 2009). These studies paint a somewhat naïve and overoptimistic picture, which overlooks the fundamental differences between internet and corporate settings. While the transparent nature of wikis may well suit the open internet settings, it may not be appropriate in settings where users are driven by career advancement and accountability is essential (Patterson et al. 2007; Arazy & Stroulia, 2009).

The objective of this paper is to study the life cycle of corporate wiki adoption. Users' wiki activity (i.e. modifications of the wiki page or simply 'edits') is automatically logged by the wiki engine, and this paper we investigate the temporal edit patterns of corporate wikis. While a number of studies describe the activity patterns of Wikipedia (Viegas et al. 2004), we are not aware of any prior works that studied the life cycle and temporal activity patterns of corporate wikis. Open source software development is in many ways similar to wiki-based collaboration (Wagner 2006), and in a study of this related area Crowston et al. (2006) have identified several prototypical temporal activity patterns, including "consistency rising teams" and "consistently falling team". The aim of this paper is to characterize - both quantitatively and through visualizations - wiki activity patterns, and explore whether these patterns resemble the patterns observed for open source software development. We expect that our analysis would provide a glimpse

into wiki-enabled collaboration processes. Those developing theories for wiki work processes could employ our results to develop hypothesis regarding the factors that drive wiki activity.

2. Related Work

Very little is known about corporate wiki adoption. Some evidence regarding wiki adoption is provided in studies that surveyed wiki users (e.g., Patterson et al. 2007). These studies shed light on the motivations for contributing to wikis and on the perceived risks, which could affect activity levels. It becomes clear that – notwithstanding their advantages – wikis are susceptible to quality threats and their transparent nature gives rise to risk-avoidance behavior that is likely to inhibit wiki activity. These findings stand in contrast to the all positive conclusion of Majchrzak et al. (2006) that corporate wikis are sustainable, and to the results of Arazy et al. (2009) that illustrated how corporate wiki growth rates surpass those experienced by Wikipedia. We expect that a large-scale investigation of temporal patters of wiki activity logs would help to resolve these discrepancies and shed light on corporate wikis adoption lifecycles.

There are various approaches for modeling and visualizing activity patterns in collaborative projects. While we are not aware of techniques that have been applied to study corporate wikis, prior studies have described activity patterns in similar contexts. Studies of Wikipedia introduced visualization that revealed the complexities of the collaborative authoring process (Viegas et al. 2004). These methods vividly describe one extremely successful wiki application. However, they are less useful for the describing the life cycle of a large number of different wiki applications. Classification and clustering techniques may be more suitable for describing alternative prototypical behavior patterns. Each collaborative project could be described as a time series, and projects could be grouped using various approaches. For example, Crowston et al. (2006) described 122 open source software development projects by the size of the developer group, and manually sorted the projects time series into six pre-defined classes: consistent risers, risers, steady or not treading, fallers, consistent fallers, and dead projects. They found evidence for all these patterns, the vast majority falling into the 'consistent risers' category. Categorization of time series could also be automated, by calculating the similarity (or distance) between any time-series pairs, and then clustering the projects based on these similarities.

Calculating the similarity between a time series pair is a key challenge. Similarities are estimated by (a) establishing correspondence between points along the two time series, (b) calculating the similarity between corresponding points, and (c) aggregating the similarities. Once the similarities between all wiki time series pairs is established, clustering could be performed using standard methods (e.g. hierarchical clustering). Under the assumption of a one-to-one correspondence between points along the two time series, a simple distance metric – e.g. Euclidian distance – could be used. This assumption, however, may not be valid when matching wiki activity time series. Since each wiki activity log represents a unique collaborative work process, each time series is expected to have a different starting points, length, and range of values. The Dynamic Time Warping (DTW) approach could be employed for estimating time series similarity under such conditions (Keogh 2004). In DTW, the overall similarity between two time series is formulated as a stepwise local optimization problem, in which non-linear one-to-many alignment is permitted, hence relaxing the one-to-one correspondence constraint. It should be noted that while warping is allowed in DTW, the temporal order of the points is preserved. Recently, the Longest Common Sub-Sequence (LCSS) algorithm (Vlachos et al. 2006) has been suggested as a further improvement of the DTW approach, which accommodates the formation of gaps (points that remain unmatched). While these methods provide generic solutions for estimating time series similarity, the unique characteristics of the problem at hand require some further enhancements. Some of these unique characteristics include: the typical 'birth-life-inactivity' project lifecycle, short temporal patterns, highly 'bumpy' patterns, and substantial variations in scale. For example, we wish to differentiate between short-lived wiki projects and those that experience a sustainable period before becoming inactive. The methods proposed by DTW and LCSS allow for unconstrained wrapping that does not distinguish between these two different temporal patterns.

3. The Proposed Method

The sample for our study consisted of all of the wikis – 13,313 distinct applications – at IBM. IBM is a global organization with over 350,000 employees that designs hardware, develops software, and engages in professional services. Wikis are used at IBM for various tasks – from use as a simple web portal to more complex applications, such as content generation (e.g., creating a product manual or FAQ database), project management, and an application to support communities of practice (Arazy et al. 2009). Our data included the monthly number of edits made to each wiki, from when the wiki infrastructure became operational (September 2005) until the cut-off date of January 2008. Analyzing and clustering these time series, and specifically estimating similarities, was a challenging task, due the distinct features of the wiki activity time series mentioned earlier. In addition, the very large size of the data set presented computational challenges. Our method included the following steps:

(a) **Generating time series.** The objective of this step was to construct, based on time-stamped edit logs, a time series of wikis lifeline. In particular, we identified three key events in each lifeline: (i) *Time of Birth*, the first month in which the wiki was launched; (ii) *Time of Inactivity* (TOI), a prolonged period of time during which no activity was recorded. We determined TOI as the beginning of the first period of consecutive 3 months during which the wiki was inactive (note that some wikis remain active until our cutoff date, and thus do not include a TOI); and (iii) *Time of End*, the last period for which data is available (the cutoff date). Thus, each wiki application was described as time series of monthly edits, from month zero (Time of Birth) until the cutoff date. It should be noted that the TOI detection method was chosen based on manual exploration of a randomly selected number of wiki time series. Changes in the definition of the TOI did not have a substantial impact on our results.

(b) **Denoting inactive wikis.** We wanted to clearly distinguish between inactive wikis and those with very low activity levels, and assigned a value of -50 to all periods after the TOI.

(c) **Addressing differences in scale**. How should wikis with very similar lifeline patterns but with different scales of activity be regarded? Our answer was that when the differences in scale are not large the wikis should be clustered together, but when activity levels are an order of magnitude apart the lifelines should be treated as different. To accomplish this, we log transformed wikis time series.

(d) **Similarity estimation.** We calculated the similarity between each pos*sible pair of wiki*s using an enhanced version of the LCSS algorithm. We constrained the occurrence and size of gaps that are formed between 'matched' points, in line with the Needleman-Wunsch algorithm (Durbin et al. 1998).

(e) **Clustering**. Based on the pair-wise similarities, a distance matrix was computed for the entire data set, and a hierarchical clustering (N=5) was applied using complete linkage.

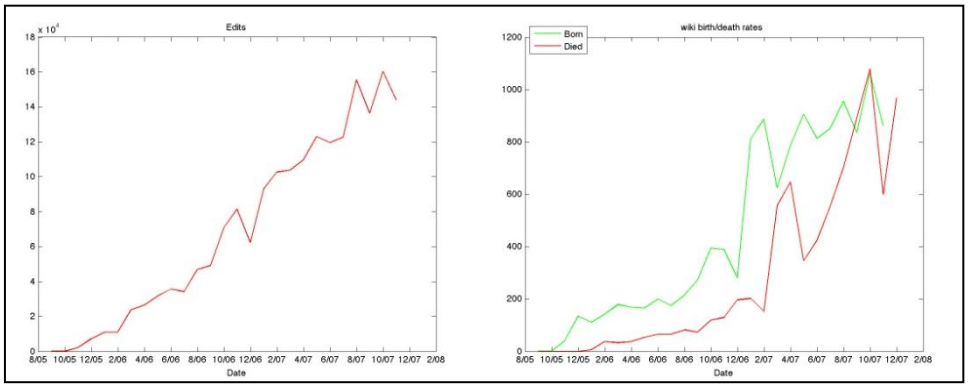
(f) **Visalization**. We generated two visualizations. First, we produced a density plot for each cluster, where we transformed wikis time series to an accumulator array, using a vector-to-raster conversion. By accumulating the number of lines passing through each array cell, we computed the overall density per period and activity range. Second, we produced a line plot for each cluster, where we assigned a color to wikis time series according to their length.

4. Results

Below we report the results for two types of analysis. First, we analyzed wiki activity over time (September 2005 – November 2007), looking at the set of all IBM wikis. We analyzed the total wiki monthly edits and compared the frequency of 'birth' and TOI events. Second, we described each wiki as a time series starting with its inception (i.e. Period 0) and classified the wikis using our proposed clustering algorithm.

Figure 1a below depicts the monthly edits. It shows a steady rise (except for a drop at holiday season) until the end of our analysis period, where activity levels start to fall. Figure 1b compares the frequency of 'birth' versus TOI events. The similarity between the two graphs is striking, illustrating how the TOI graph follows the 'birth' graph with a 2-3 month delay. This suggests that many of the wikis become

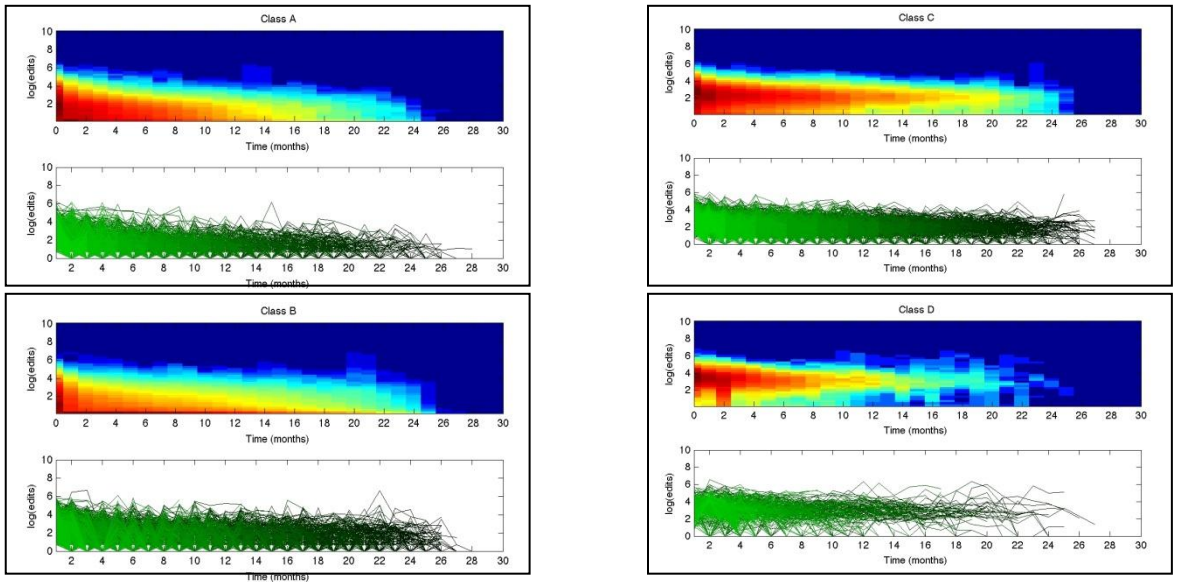
inactive shortly after their inception. We also notice that while at the initial period when wikis were introduced birth rates were rose continuously, towards the end of our analysis period TOI rates surpass 'birth' rates.



Figures 1a and 1b: The figure on the left (1a) shows total wiki monthly edits and the one on the right (1b) compares the frequency of 'birth' and TOI events.

The first observation from the analysis of wikis' lifecycles is that wikis' time series depict highly irregular patterns, with sharp rises and drops in monthly activity levels. I.e. the pattern is very 'bumpy', as illustrated in bottom of Figures 2a-2d below (in green).

The results for five high-level clusters reveal distinct lifecycle patterns. Clusters A (which included 2440 wiki applications; 18% of total wikis) and B (7679 wikis; 58%) show clusters that slowly decay until they become inactive. While both clusters start at similar activity levels, Cluster B drops in activity levels quickly, while the decay in Cluster A is more gradual. Another difference is that Cluster B maintains activity much longer than Cluster A does. Clusters C (2182 wikis; 16%) and D (381 wikis; 3%) represent wikis that are sustainable and remain active at consistent rates until the cutoff date. The main difference between these clusters is that Cluster D represents higher activity levels (roughly at 50 edits per month, versus 30 for Cluster C). Also, Cluster D initially grows in activity, while Cluster C reaches the pick at the first period and then slowly drops in activity levels. Cluster E (631 wikis; 5%; not presented in the figures) represent wikis that did not reach maturity and were artificially truncated because of the cutoff date, and thus is an artifact of our data.



Figures 2a-2d: temporal activity patters for the primary four clusters. The X axis shows the periods since wikis' 'birth' and the Y axis shows the number of edits (log transformed). For each cluster, the top graph represents the density in number of wikis at each activity level (highest density in red; lowest in blue). The bottom graph shows the time series for all wikis in that cluster.

5. Discussion and Conclusion

Prior research on wikis' affordances (e.g. Wagner 2006) was based on an investigation Wikipedia, and proposed that corporations could adopt Wikipedia-like processes to alleviate knowledge acquisition bottlenecks. This naïve view of wikis' capabilities was supported by recent surveys of corporate wiki adoption (Majchrzak et al. 2006; Arazy et al. 2009). However, the clear disparity between volunteer based self-governed Wikipedia and traditional command-and-control corporate governance suggests that wikis may not be suitable for all organizational contexts. Are wikis, then, suitable, for corporate settings? To date, little is known regarding the actual adoption life cycles of corporate wikis.

In this paper we've tried to address this gap by proposing a novel clustering method for categorizing temporal activity patterns of wiki edits. Existing approaches for estimating the similarity of time series (i.e. DTW and LCSS) allow matching series of varying lengths by wrapping. However, the problem at hand presented some unique challenges. In order to cluster wikis' temporal activity time series we: determined clear 'birth' and inactivity events, log-transformed the data, and constrained the wrapping to distinguish between varying levels of activity decays. In order to visualize the differences between wiki lifecycle clusters, we used both wiki lifeline plots and cluster density diagrams.

The principal findings from our analysis are that the majority of wiki applications are not sustainable over a long time period, as opposed to what has been suggested in prior survey-based studies (e.g. Majchrzak et al. 2006). We believe that the exponential growth in overall activity levels that were reported in prior studies (e.g. Arazy et al. 2009) stem from the early hype period. However, as wikis are reaching maturity, we observe that many applications become inactive. Towards the end of our analysis period the number of total wiki monthly edits begins dropping, and TOI rates surpass 'birth' rates. The delay between the 'birth' and TOI graphs suggests that often users experiment with the new technology and then soon abandon it. Contrary to our expectation that wikis would exhibit relatively stable activity patterns, the activity time series were extremely 'bumpy'. Our clustering analysis revealed four primary lifecycle patterns: 'fast plummet' (Cluster A), 'slow plummet' (B), 'constantly weakly-active' (C), and 'constantly highly-active' (Cluster D). The plummeting clusters (A and B) captured over 75% of the 13,313 wikis at IBM, while the clusters with continuous activity (C and D) included less that 20% of the total wikis, demonstrating that the majority of wiki application are active only for short periods. Interestingly enough, these patterns are quite different from the temporal patterns reported for open source software projects (Crowston et al. 2006). The differences may stem from a number of reasons (e.g., the type of project, underlying technology, or the organizational setting), and could be explored in future research.

The primary contributions of this paper are in (i) enhancing our understanding of corporate wiki life cycles and (ii) the extensions made to the method for estimating time series similarity. Our analysis revealed some novel findings that stand in contrast to the results reported in earlier studies. Future work is warranted in order to: enhance the time series clustering and visualization methods, analyze wiki lifecycles over longer time periods, explain the discrepancies from previous results, and extend the analysis to other settings. Specifically, in the future we plan to investigate what happens after wikis become inactive (has the wiki-based project failed, was the wiki's purpose served, are users still accessing the wiki such that the wiki is impacting organizational learning after becoming inactive), by analyzing wiki page visits. In addition, we plan to explore the characteristics of wikis in each of the clusters (and differences between clusters) by surveying the users of these wiki applications (e.g., are the motivations of users in sustainable wikis differ from the motivations of the plummeting wikis?). In conclusion, wiki is a promising technology that has the potential to transform knowledge management. However, much research is needed for determining the specific situations in which such a decentralized collaborative technology could succeed in corporate settings. Those developing theories for wiki work processes could employ our results to develop hypothesis regarding the factors that drive wiki activity.

Acknowledgements

This research was funded in part by SSHRC and NSERC.

References

Arazy O., Gellatly I., Jang S., and Patterson R.., "Wiki Deployment in Corporate Settings", *IEEE Technology and Society*, Volume 28, Number 2, Summer 2009, pp. 57-64.

Arazy O. and Stroulia E., "A Tool for Estimating the Relative Contributions of Wiki Authors", in *Proceedings of ICWSM'09*, May 2009, San Jose, California, USA.

Crowston, K., Howison, J., and Annabi, H., "Information systems success in free and open source software development", *Software Process: Improvement and Practice*, 2006, 11:2, pp. 123-148.

Durbin R., Eddy S., Krogh A. and Mitchison, G., "Biological sequence analysis". 1998, Cambridge University Press.

Keogh, E., "Exact indexing of Dynamic Time Warping". In *Proceedings of the 28th VLDB Conference*, Hong Kong, China, 2002, pp. 406-417.

Majchrzak A., Wagner C., and Yates D. "Corporate wiki users: results of a survey," in *Proceedings of the international symposium on Symposium on Wikis*, 2006, pp. 99-104, ACM Press.

Majchrzak, A. "Comment: Where is the theory in wikis?" *MIS Quarterly*, 2009, 33 (1), pp. 18-20.

Patterson R., Gellatly I., Arazy O., and Jang S., "The Effects of Wikis Characteristics on Performance Quality", in *Proceeding of WITS'07*, December 2007, Montreal, Canada.

Viégas, F. B., Wattenberg, M., & Dave, K., "Studying Cooperation and Conflict between Authors with history flow Visualizations". In *Proceedings of CHI'2004*, 2004, Austria, pp. 575-582.

Vlachos, M., Hadjieleftheriou M., Gunopulus D. and Keogh, E., "Indexing multidimensional time series". *The VLDB Journal, 2006,* 15(1), pp.1-20.

Wagner, C., "Wiki: A technology for conversational knowledge management and group collaboration." *Communication of the Association for Information Systems*, 2004, 13, pp. 265-289.

Wagner C., "Breaking the knowledge acquisition bottleneck through conversational knowledge management," *Information Resources Management Journal*, 2006, 19 (1), pp. 70–83.

DO WIKI-PAGES HAVE PARENTS? AN ARTICLE-LEVEL INQUIRY INTO WIKIPEDIA'S INEQUALITIES

Abhishek Nagaraj, Amitava Dutta, Priya Seetharaman, Rahul Roy
Indian Institute of Management Calcutta, George Mason University, Indian Institute of Management Calcutta, Indian Institute of Management Calcutta,
abhishekn2010@email.iimcal.ac.in, adutta@gmu.edu, priyas@iimcal.ac.in, rahul@iimcal.ac.in

Abstract

We hypothesize that articles on Wikipedia have "parents" who contribute a significant portion of their edits. We establish a notion of inequality based on the Gini Co-efficient for articles on Wikipedia and find support for the existence of this phenomenon of parenting. We base our study on data collected from the Tagalog and Croatian Wikipedias. Ultimately we claim that our research has significant implications for policy for both Corporate Wikis as also for Wikipedia. We state these implications and also suggest directions for future research.

Keywords: Wikipedia, Inequality, Parenting, Knowledge Management, Corporate Wikis

1. Introduction

Wikipedia the online collaborative encyclopedia has captured the attention of not only scholars from a variety of fields but also from mainstream media. One of the fundamental objectives of these investigations has been to determine the reasons for Wikipedia's ability to nearly match other respected publications such as the Encyclopedia Britannica in terms of article quality (Giles 2005). A variety of parameters based on page characteristics have been used to explain differences in article quality. These range from simple parameters like word count (Blumenstock 2008) to more complex models linking article quality to author authority and peer reviews (Hu et al. 2007).

Another important line of investigation has been to look at contributors themselves and explain their behavior. At the very basic level authors have been classified based on simple properties like edit counts and the period for which they have been active. A study by Kittur et al. (2007) for example uses this distinction to examine the changing influence of "elite" and "common" users over time in Wikipedia. An important study in this category has been the one by Anthony et al. (2005) which contends that two types of users contribute significantly to article quality – the "Good Samaritans", one time users who make high quality contributions and the "Zealots", committed users who have been contributing significantly over the past.

The present study lies primarily in that class of papers which tries to identify a particular category of contributors and links them to article quality. We call this category "parents". In the following sections we define what we mean by "parents" and list a few of the characteristics that parents demonstrate. We then describe our methodology and use inequality measures to find support for the phenomenon of "parenting". In the concluding section we make suggestions about the implications of such a finding and directions for future research.

2. Objectives and Hypothesis

In order to look for evidence of parenting in Wikipedia we draw from economics literature to apply the concept of the "Gini Coefficient" introduced by Corrado Gini to measure the inequality of wealth distribution in a population (Gini 1936). We use this parameter to define and measure the inequality in contributions for a particular article. We define the number of contributions made by a particular contributor to a particular article as his "wealth" and the total number of contributions to a particular article as the "total wealth" of that particular article. We apply the Gini Coefficient defined in this manner

to calculate inequality over a particular article and contend that a high degree of inequality for a given article signifies that it is being "parented" by a few users.

To calculate the Gini Coefficient we first plot the Lorenz curve for a particular article. The Lorenz curve is defined as "a graphical representation of the cumulative sum of contributions where we sort contributors on the horizontal axis by their amount of contribution" (Ortega et al. 2008). For a population with zero inequality i.e. one where all contributors have an equal number of edits, the Lorenz curve is a straight line – the diagonal in a unit square with side equal to the total number of contributions. For any other value of user contributions, the Lorenz curve will be convex and will lie below the imaginary line of perfect equality. The area between these two figures is the Gini co-efficient. Thus for a perfectly unequal situation (i.e. where one user makes all the contributions) the Gini co-efficient is one while in the cases of perfect equality it is zero. In other cases it will lie between these two values.

The Gini Coefficient has so far been rarely applied to look at user contributions on Wikipedia. A recent study by Ortega et al (Ortega et al. 2008) investigated over 10 different editions of Wikipedia looking at the inequality in the distribution of the sum total of contributions for each edition. They find that Wikipedia as a whole demonstrates a large degree of inequality which remains stable over time. There is however nothing to be said about the differences in inequality among different articles and the inequality in contributions for a given article, both factors important to make conclusions about the existence of the notion of parenting.

The contribution of this paper is to use this inequality effect to look at article level inequality to find support for the phenomenon of parenting.

3. Parenting in Wikipedia

In this section we shall describe our methodology and describe our results. We conducted our studies in two stages. We first used the Tagalog[1] Wikipedia, a small-medium sized Wikipedia for our studies. Once we were reasonably sure of our claims we conducted further analyses on the Croatian Wikipedia, a much larger edition. Our choice of Wikipedia editions was based on a variety of factors including the total number of articles, the total number of edits, the total number of users, the total number of "active users" and the "depth" of the edition. A latest estimate of these figures and their definition can be obtained via the Wikimedia foundation[2].

The entire dump of these versions of Wikipedia was downloaded as on 30th July 2009. The dump is provided by the Wikimedia foundation[3] and it lists each article and the history of edits made by all users to each article along with other details like a timestamp of the edit and the username or the IP address of the user if he is not registered. Once this dump was obtained we cleaned it to remove entries made by bots. Bots are automated programs which troll Wikipedia performing a variety of functions like adding missing reference sections and reverting vandalism. We also deleted non-article pages like discussion pages or categorization pages. Further analyses were performed on these cleaned versions.

The Tagalog Wikipedia contained 29089 unique pages and 13859 unique user ids. While there were ~420k total revisions this number reduced to ~120k after the data was cleaned. This formed the dataset for our analyses. On initial analyses we were able to verify a well known fact about Wikipedia – most users would contribute only one edit. This is shown in Figure 1 where we plot the number of contributors on the Y axis and frequency on the X axis.

[1] Tagalog is a language mainly spoken in the Philippines

[2] http://meta.wikimedia.org/wiki/List_of_Wikipedias

[3] http://download.wikimedia.org/hrwiki

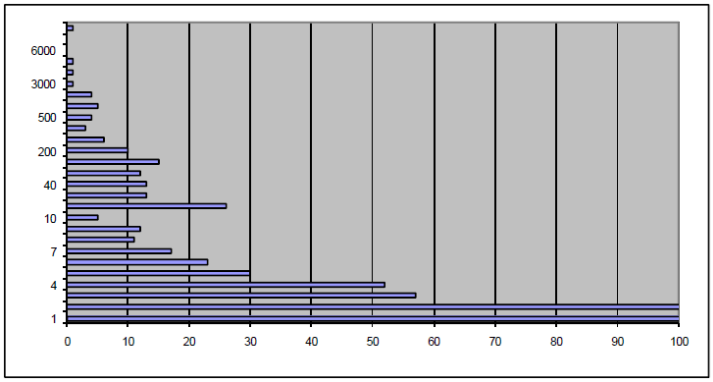


Figure 1. Frequency Distribution of User Contributions

Apart from this, the Tagalog Wikipedia also contains a large number of articles which have only one edit. 80.99% of the articles have two edits or less. This is shown in Figure 2.

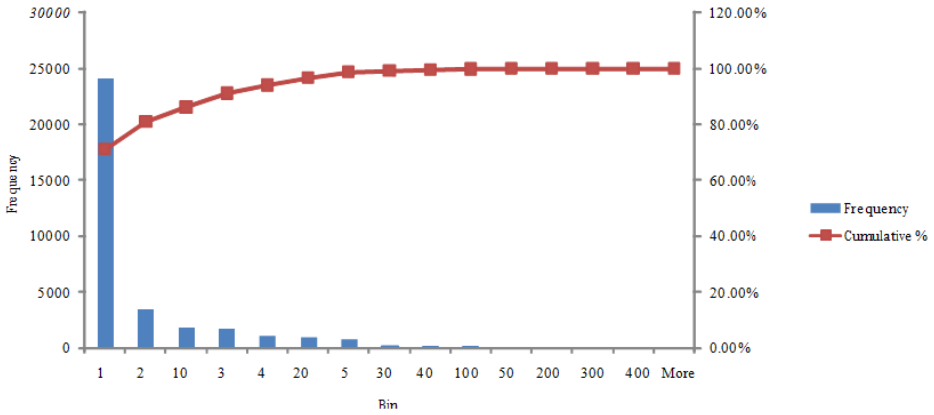


Figure 2. Histogram of Number of Article Edits

Once we had established this overall level of inequality in the Tagalog Wikipedia we then turned to looking for traces of parenting. As an initial investigation we looked at the top 4 articles by edits on the Tagalog Wikipedia. These are as evidenced by their titles the most popular articles on this Wikipedia. Topics like the country of the Wikipedia "Philippines" and its capital "Manila" are bound to attract editor attention. Yet as shown in Table 1 parents are able to capture these pages and contribute to them in a significant way. This gives us a starting point to trace such parenting features in a larger Wikipedia on a more systematic basis.

Table 1. Top Pages by Total Edits in the Tagalog Wikipedia

Rank	Page Title	Parent	Parent Edits	Total Edits	% by Parent
1.	Unaang	Kampfgruppe	288	297	96.97%
2.	Maynila	DragosteaDinTei	165	234	70.51%
3.	talaan mga bansa	AnakngAraw	201	231	87.01%
4.	Pilipinas	Bluemask	79	196	40.31%

In order to not restrict our investigation to finding the "top parent" we turn to the Gini Coefficient as described above. By using "inequality" as a measure of parenting instead of simply the top user by edits as shown above we are now able to capture a variety of cases where a few users might parent a single page or where parenting as a feature is simply absent. We use this newly introduced measure to look at the Croatian Wikipedia. The Croatian Wikipedia after cleaning was found to contain approximately 1.5M user revisions and about 138k unique articles.

We now turn to measuring the inequality of pages in this Wikipedia. Wikipedia defines "Featured Articles" to be the best quality articles on Wikipedia[4]. This classification is based on decisions taken by contributors and must satisfy a stringent list of criteria[5]. We use "Featured Articles" to be a convenient proxy for "high quality articles" and use the list of Featured Articles on the Croatian Wikipedia for further analyses. As of July 30, 2009 there were 223 Featured Articles on the Croatian Wikipedia. We sampled 25 articles out of this list randomly. We also sampled a list of 25 non-featured articles randomly controlling for mean number of article edits in the second case.

Now in order to calculate the Gini Co-efficient for these two sets we used the formula proposed by Angus Deaton (1997) as given in Figure 3.

$$G = \frac{N+1}{N-1} - \frac{2}{N(N-1)u}(\Sigma_{i=1}^{n} P_i X_i)$$

Figure 3. Formula to calculate the Gini Coefficient

In the above formula, G represents the Gini Coefficient for a particular page, N represents the total number of edits for a particular article, u represents the mean number of edits, P_i represents the rank of a particular contributor where Rank 1 is held by the "richest" contributor in terms of edits and X_i represents the total number of edits or the "wealth" of a particular contributor. Using the above formula we are now in a position to calculate the Gini Co-efficient for the two sets of randomly selected 25 articles. These articles have an average of about 183 edits per article. The results are shown in Table 2 and Table 3. Figure 4 shows how a graphical representation of this calculation for the article "Zagreb".

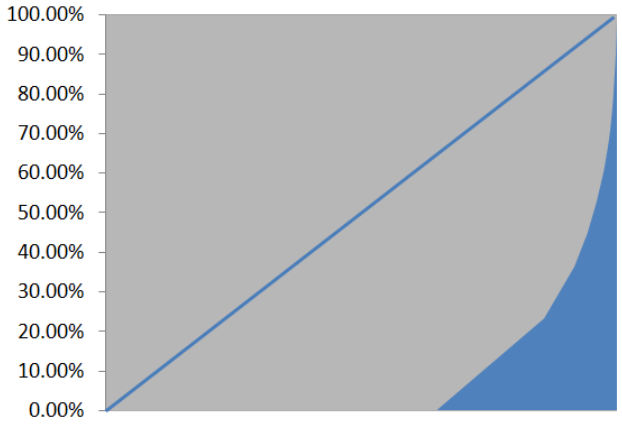


Figure 4. Lorenz Curve for the article "Zagreb".
The area between the line and the curve is the Gini Co-efficient, in this case 67.44%

[4] http://en.wikipedia.org/wiki/Featured_Article

[5] http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

Table 2. Featured Articles	
Page Title	**Gini Coefficient**
Arnold Schanberg	72.86%
Asirija	77.01%
Autizam	73.36%
Dioniz	59.13%
Filozofija	68.75%
Francisco Franco	68.08%
Gabriel Garcia Marquez	60.73%
Gospodar prstenova	65.97%
Gospodarstvo Bocvane	85.97%
Harry Potter i Darovi smrti	66.82%
Holokaust	59.58%
Hrvatski jezik	70.27%
Indijanci	88.99%
Jean-Paul Sartre	67.97%
Jezik	65.89%
Nizozemski jezik	66.74%
Nordijska mitologija	81.46%
Odisej	66.85%
Orgazam	72.43%
Sherlock Holmes	60.98%
Staroslavenski jezik	67.65%
Ukrajinci	78.14%
Vikinzi	61.21%
William Shakespeare	60.75%
Zagreb	67.44%

Table 3. Non-Featured Articles	
Page Title	**Gini Coefficient**
Adolf Hitler	68.22%
Borema (nogomet)	74.71%
Bosna i Hercegovina	62.65%
Britney Spears	67.02%
Crna Gora	62.56%
Donji Miholjac	59.07%
Grude	61.56%
HNK Hajduk Split	81.99%
Hrvati	67.63%
Hrvatska nogometna reprezentacija	72.31%
Hrvatska Republika Herceg-Bosna	78.89%
Hrvatski demo sastavi	65.50%
Kosovo	65.28%
Livno	65.80%
Nezavisna Drint	63.75%
NK Dinamo Zagreb	79.10%
Nordijska mitologija	65.43%
Osijek	67.76%
Predlo	74.06%
Rijeka	66.78%
RNK Split	76.08%
Slavonski Brod	65.95%
Slovenija	59.19%
Split	69.47%
Srbi	68.17%
Srbija	62.11%

The average Gini Co-efficient for Featured Articles is 61.44% (std. deviation 7.61%) while that for Non-Featured articles is 68.12% (std. deviation 6.15%). This finding strongly suggests that there is a high degree of inequality in Wiki-pages; that is there is strong evidence for the presence of a small group of users who "parent" articles. For our finding we have used pages of high quality (Featured Articles) and pages with a high number of revisions. These are pages that are in some sense "popular" and would intuitively be the hardest for a particular group of "parents" to dominate. Yet, we see that this is exactly

what happens. We find support for the proposition that even in Wikipedia's most popular parts parenting exists. This is our contribution to the existing understanding of Wikipedia.

4. Implications and suggestions for future research

The major implication of our study is the suggestion that for well-established mature pages to develop the phenomenon of parenting is inevitable. Pages develop and mature when they find a few people ready to nurture it. This leads to two concrete suggestions, one for the business world and the other for Wikipedia administrators. For corporate wikis there is undoubtedly a case to be made for an explicit allotment of people to pages, hoping that such attention would cause the pages to mature. Wikipedia policy on the other hand should look at easing this process of pages finding parents either by measures like explicitly creating "parent" roles or "become-a-parent" suggestion boxes based on edit history.

As for future research, there is obviously the need to further investigate the importance of parenting in featured pages achieving their status. This would strengthen the idea that parenting leads to higher quality. Our study also opens up the possibility of using hybrid parameters to test page quality which include the Gini Coefficient. Another interesting thing to do would be to conduct this study in an orthogonal manner i.e. look at the inequalities among user contributions across pages. That could further strengthen our hypothesis of contributors being parents of some articles and fleeting editors on others. Lastly it would be interesting to test for a notion of "good" and "bad" parenting – parents who nurture pages and moderate discussion as opposed to parents who impose their view on an article. Tying in this perspective of parenting to previous literature like Anthony et al. (2005) could lead to interesting results.

By proposing and finding support for this phenomenon of parenting we have thus opened up a new perspective which could lead to interesting results in the future.

References

Anthony D., Smith S. and Williamson, T., "Explaining quality in internet collective goods: Zealots and good samaritans in the case of Wikipedia", November 2005. Retrieved online. *http://web.mit.edu/iandeseminar/Papers/Fall2005/anthony.pdf*.

Blumenstock, J.E. "Size matters: word count as a measure of quality on wikipedia", in *Proceedings of the 17th international conference on World Wide Web*, ACM, New York, USA (Pub.), April 2008, pp. 1095-1096.

Deaton A, "The analysis of household surveys: a microeconometric approach to Development Policy", The John Hopkins University Press, Baltimore, 1997, pp. 137.

Giles, J. "Special Report: Internet encyclopaedias go head to head", *Nature* 438, December 2005, pp. 900-901.

Gini C., "On the Measure of Concentration with Special Reference to Income and Wealth", in *Abstracts of Papers Presented at the Cowles Commission Research Conference on Economics and Statistics*, Colorado College Publications, 1936.

Hu M., Lim E., Sun A., Lauw H.W. and Vuong B., "Measuring article quality in wikipedia: models and evaluation", in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management,* ACM, New York, USA (Pub.), November 2007, pp. 243-252.

Kittur A., Chi E. H., Pendleton B. A., Suh B., Mytkowicz T., "Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisi", in *Proceedings of the 25th Annual ACM Conference on Human Factors in Computing Systems,* ACM, New York, USA (Pub.), April 2007.

Ortega F., Gonzalez-Barahona J.M., Robles G., "On the Inequality of Contributions to Wikipedia", in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences,* IEEE Computer Society Press, Big Island, Hawaii, January 2008, pp. 308.

Who Does What: Collaboration Patterns in the Wikipedia and Their Impact on Data Quality

Jun Liu, Sudha Ram
Department of Management Information Systems, Eller College of Management,
The University of Arizona, Tucson, AZ 85721, USA

Abstract

Data quality in the Wikipedia is debatable. On the one hand, existing research indicates that not only are people willing to contribute articles but the quality of those articles is close to that found in conventional encyclopedias. On the other hand, the public has never stopped criticizing the quality of Wikipedia articles, and critics never have trouble finding low quality Wikipedia articles. Why do Wikipedia articles vary widely in quality? We investigate the relationship between collaboration and data quality. We show that the quality of Wikipedia articles is not only dependent on the different types of contributors but also on how they collaborate. Based on an empirical study, we classify contributors based on their roles in editing individual Wikipedia articles. We identify various patterns of collaboration based on the provenance or, more specifically, who does what to Wikipedia articles. Our research helps identify collaboration patterns that are preferable or detrimental for data quality, thus providing insights for improving data quality in Wikipedia.

Keywords: Wikipedia, collaboration pattern, data quality, data provenance

1. Introduction

There has been an interesting debate lately about the quality of Wikipedia, the free online encyclopedia. Many believe the quality of Wikipedia articles to be surprisingly good despite its seemingly bizarre everyone-can-edit principle. A much discussed article from Nature (Giles 2005) compares Wikipedia with the Britannica Encyclopedia and argues that despite its anarchical operation, the former comes close to the latter in terms of the accuracy of its science entries. Nevertheless, critics keep attacking Wikipedia since "no one stands officially behind the authenticity and accuracy of any information in Wikipedia" (Denning et al. 2005). After all, only 2,587 out of a total of 2,994,903 articles on the English Wikipedia are slated to be featured articles - articles that are "professional, outstanding and thorough" (Wikipedia 2009) .

Why are some Wikipedia articles of high quality while others are not? Most of the current research considers collaborations as a critical reason for high-quality Wikipedia articles. Lih (2004) suggests metrics such as "rigor" (total number of edits made for the article) and "diversity" (total number of unique editors for the article) as measures of quality. An article with a large number of edits and editors is often of high quality since "given enough eyeballs all bugs are shallow" (Lih 2004). However, research such as (Lih 2004) does not consider the diversity of editors and their contributions. The fact that Wikipedia is easy to edit does not mean that all contributors edit the same way, or with the same intensity. Anthony et al (2005) took a different approach with their study. They believe that quality depends entirely on the types of contributors to Wikipedia. High quality content has been shown to come from two types of users – *zealots*, registered users with a strong interest in reputation and high level of participation and *good Samaritans*, unregistered, anonymous and occasional contributors.

Drawing upon existing research such as (Lih 2004; Anthony et al. 2005), our research attempts to investigate the relationship between collaboration and data quality in Wikipedia. We believe the quality of Wikipedia articles is not only dependent on the different types of contributors but also on how they collaborate.

2. Overview of our research

Figure 1 shows the theoretical model that forms the foundation for this study. Consistent with the McGrath framework (McGrath 1984) that has been adopted by many collaboration researchers, we use an input-process-output framework for identifying the key components in our research.

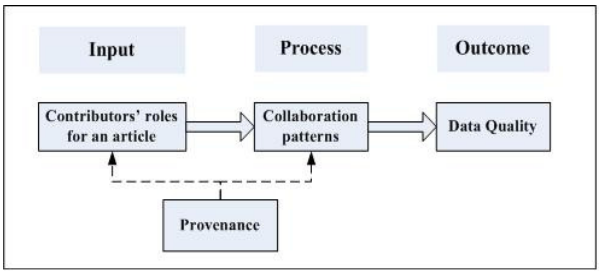


Figure 1. Theoretical model of our research

Each Wikipedia article has contributions from many different individuals. The number and types of contributions are known to vary from individual to individual since each contributor is motivated differently and has a different level of familiarity with the topic. In our research, we first identify various roles played by contributors for a given article. A role consists of a set of expected and enacted behaviors of a contributor. Using these roles we uncover collaboration patterns. Each of the patterns represents a distinctive way in which a group of contributors who play different roles collaborate. We identify the roles of contributors and collaboration patterns based on the provenance (defined in the next section) of each Wikipedia article. We then examine the quality of the articles to determine the impact of collaboration patterns on quality of the Wikipedia articles. In the rest of this paper, we describe each of the components of our research in more detail.

3. Data provenance of Wikipedia articles

Data provenance refers to the source and processing history of data. We track and use the provenance of Wikipedia articles. Ram and Liu (2007) have clearly defined the concept of provenance using the W7 model. We employ a subset of this model by tracking every action that affects the life of a Wikipedia article from its creation to the present time. We also use the information about the specific contributor who performed each action, and the time the action occurred in the life of the Wikipedia article. There are many actions that can affect a Wikipedia article as shown in Table 1. A contributor makes an edit to a Wikipedia article by performing one or more actions.

Table 1: Definition of actions that can affect a Wikipedia article

Type of actions	*Explanation*
Sentence creation	Creation of a sentence
Sentence modification	Modification or rewording of an existing sentence
Sentence deletion	Deletion of a sentence
Link creation	Linking of a word within an existing sentence to a article (a link to another Wikipedia article or to external Internet articles)
Link modification	Modification of an existing link (can be a change of the URL or the name of the link)
Link deletion	Deletion of an existing link
Reference creation	Adding a reference or creation of an inline citation
Reference modification	Modification of an existing reference
Reference deletion	Deletion of a reference
Revert	Reverting a article to a former version

4. Identification of contributor roles in the Wikipedia

Identifying the roles played by each contributor helps us understand the sources of quality variance in the Wikipedia. Extant research such as (Bryant et al. 2005) investigates contributors' roles in the Wikipedia community. However, a contributor's role may vary from one article to another. We focus on contributors'

roles specific to an article. We cluster users of each article based on their contributions to that specific article. To do this, we employ the K-means clustering technique.

1. *Inputs to clustering:* If we use *P* to represent a set of Wikipedia articles and *cp* to represent a contributor who has contributed to a article $p \in P$, then a sample going into the clustering can be represented as a vector $\overrightarrow{cp} = <E_1^{cp}/E_t^{cp},\ E_2^{cp}/E_t^{cp} \ldots,\ E_{11}^{cp}/E_t^{cp},\ D^{cp}>$, where $E_1^{cp},\ E_2^{cp},\ ..,\ E_{11}^{cp}$ represent the number of each of the 10 types of actions (see Table 1) performed by the contributor *cp* to a given article *p*. We do not include actions that were immediately reverted or deleted. E_t^{cp} represents the total number of actions performed by the contributor to the article, and D^{cp} the number of days the contributor edited the article, which is derived from the time associated with each action.

2. *Data collection:* The data set we used in this study consists of articles collected from the English Wikipedia in June 2009. We took advantage of Wikipedia's article assessment project, which has organized the evaluation over 900,000 articles into various grades of quality ranging from "featured article" to "C-class article" status. We randomly collected 1600 articles including 400 featured articles, 400 A-class articles, 400 B-class articles and 400 C-class article as our data set. As described above, each vector used in clustering represents the behavior of a contributor on a given article. The randomly collected 1600 articles contain a total of 1636801 such vectors. We also noticed that 90.78% of contributors had less than 4 actions for a given article. We categorized these contributors as *casual contributors* for the article and did not include them in clustering. As a result, the data used for clustering include 163576 vectors.

3. *Repeated K-means algorithm:* We used the well-known K-means algorithm as the base method to cluster the sample data. A well-known disadvantage of K-means is that it requires the number of clusters, *k*, to be specified a priori. To address this problem, we followed (Liu and Keselj 2007) and applied the K-means method repeatedly using *k* values ranging from 2 to 10. Here, we set 10 as the maximum k value to avoid a trivial classification of roles. For each *k* value, we first evaluated the quality of the clustering results using evaluation functions proposed by (Niu et al. 2006), i.e., *cluster compactness (Cmp)*, *cluster separation (Sep)* and *combined measure of overall cluster quality (Ocq),* to evaluate both the intra-cluster homogeneity and inter-cluster separation of the clustering result. The definitions of these functions are given below.

$Cmp = \frac{1}{C}\sum_i^C \frac{v(c_i)}{v(X)}$, where *C* is the number of clusters generated on the data set *X, v(c_i)* is the deviation of the cluster c_i and *v(X)* is the deviation of the data set *X*. $v(X) = \sqrt{\frac{1}{N}\sum_{i=1}^N d^2(x_i, \bar{x})}$, where *d( )* is a distance measure between two vectors *N* is the number of members in *X,* and $\bar{x}$ is the mean of *X*.

$Sep = \frac{1}{C(C-1)}\sum_{i=1}^C \sum_{j=1, j\neq i}^C \exp\left(-\frac{d^2(x_{c_i}, x_{c_j})}{2\sigma^2}\right)$, where σ is the standard deviation of the data set *X*, *C* is the number of clusters. x_{c_i} is the centroid of the cluster c_i, $d(x_{c_i}, x_{c_j})$ is the distance between the centroid of c_i and the centroid of c_j.

$Ocq = 0.5 \times Cmp + 0.5 \times Sep$. The lower the *Ocq* value, the better the quality of the overall output clusters.

In our study, 6 was the optimal number of clusters generated from the data set because *Ocq* had the lowest value at k = 6. Table 2 shows a summary of the 6 clusters that were generated. We assigned a role label to each of these clusters to designate the role played by the contributors. We categorized the contributors that belong to cluster 1 as *all-round contributors* since they were engaged in almost all types of actions. Contributors in cluster 2 were labeled as *watchdogs* since most of their actions were reverts. Cluster 3 included contributors who created sentences while seldom engaging in other actions and were hence called *starters*.

Contributors that belonged to cluster 4, on the other hand, not only created sentences, but justified them with links and references. They were therefore classified as *content justifiers*. Both *starters* and *content justifiers,* however, rarely modified existing sentences created by themselves or other people. Cluster 4 included *copy editors* who contributed primarily through modifying existing sentences. Finally, those who primarily focused on removing incorrect sentences, references and links were termed *cleaners*. Thus, a contributor for a given Wikipedia article could assume one of these 6 roles or could be a *casual contributor*.

Table 2. Summary of contributors

Cluster #	*Description of actions by contributors*	*Role Label*
1	Engaging in many types of actions including sentence creations, modifications, and deletions and link and reference creations, modifications and deletions. Performing actions more frequently than an average contributor	All-round Editors
2	Focusing on reverts. Performing actions more frequently than an average contributor	Watchdogs
3	Focusing on sentence creations and seldom engaging in other actions. Performing actions less frequently	Starters
4	Focusing on three types of actions: sentence creations, link creations and reference creations. Performing actions less frequently	Content Justifiers
5	Focusing on sentence modifications	Copy Editors
6	Focusing on removing sentences, references and links	Cleaners

5. Identification of collaboration patterns

As the next step, we wanted to investigate how contributors assuming different roles collaborate with each other for each Wikipedia article. For instance, there may be Wikipedia articles where starters create a large chunk of text and then casual contributors are relied upon to modify it; or articles where all-round contributors form a core group that insert much of the content and then continuously modify their own and other people's insertions. We attempted to identify collaboration patterns among the contributors with different roles. We used clustering to group Wikipedia articles based on roles and actions performed by contributors in these articles.

We used the 1600 randomly selected Wikipedia articles described in Section IV as the data set. We identified collaboration patterns by examining *who* does *what* for these articles. The collaboration among contributors with different roles for the article *p* is represented as a vector $\vec{p} = < R_{ij}^{p}/E_{i}^{p} >, i = 1..11,\ j = 1..7$, where E_{i}^{p}, $i = 1..11$, represents the total number of one type of action (e.g., sentence creation) that occurred to the article, and R_{ij}^{p}, $j = 1..7$, the total number of one type of action (e.g., sentence creation) performed by one type of contributor (e.g. all-round contributors) to the article. We constructed the vectors for all of the selected 1600 articles and use them as input to the repeated K-means clustering algorithm described in section IV. We set *k*, the number of clusters, to vary from 2 to 10. The repeated K-means algorithm resulted in 5 as the optimal number of clusters. Table 3 shows the characteristics of the 5 clusters, and each cluster is a set of articles and has a corresponding collaboration pattern.

Table 3.Description of clusters and their corresponding collaboration patterns

Cluster#	*Collaboration pattern description*
1	Content justifiers dominated in sentence creations (account for, on average, 72% of sentence creations), reference creations (67%), and link creations (77%). Casual contributors played an important role in sentence, link and reference modifications.
2	All-round contributors conducted 44% of sentence creations, 40% of sentence modifications, 47% of sentence deletions and 70% of reference creations, 51% of reference deletions, 36% of link creations, 34% of link modifications, and 41 % of link deletions. Starters also perform 23% of sentence creations.
3	Compared with other clusters, casual contributors played a more important role. Casual contributors

	contributed 48% of sentence creations and 56% of sentence modifications. They also created many references and made reference modifications (58% and 50% respectively). Cleaners carried out 58% of sentence deletions and 51% of link deletions.
4	All-round contributors dominated. They made 75% of sentence creation, 58% of sentence modifications, 74% of deletions, 90% of reference creations, 69% of link creation and 63% of link deletions. In addition, copy editors made 24% of sentence modifications.
5	Starters dominate sentence creations (53%). Causal contributors played important roles in reference/link creations and modifications and they are also responsible for 44% of sentence modifications. Meanwhile, copy editors also contributing 24% of sentence modifications.

Note: Reverts and watchdogs were not included in the pattern description since watchdogs performed most of the reverts (at least 78%) for pages in all of the clusters.

6. Relationship between collaboration patterns and data quality

Next, we examined the quality of articles in each cluster described in table 3. We examined the correlation between the Wikipedia designated quality value and the collaboration pattern for each article in each cluster. The collaboration patterns and data quality are strongly correlated (Kendall's Tau-c = .15, p <.001) as shown in Table 4. For instance, articles that belong to cluster 4(where all-round editors dominated) are of high quality with 50% of them being designated as featured articles and 42% as A-class articles by the Wikipedia. The quality of articles in cluster 3 (where casual contributors played a dominant role) and in cluster 5 (where starters dominated sentence creations), on the other hand, is often questionable.

Table 4. Relationship between collaboration patterns and data quality [a]

	Cluster 1		*Cluster 2*		*Cluster 3*		*Cluster 4*		*Cluster 5*	
Amount	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
Featured	17	8	62	23	4	2	315	50	2	1
A-class	48	21	71	26	10	4	265	42	6	3
B-class	58	26	93	34	119	46	20	3	110	50
C-class	102	45	47	17	123	48	26	4	102	46
Total	225	100	273	100	256	100	626	100	220	100

a. Kendall's Tau-c = .15, p < 0.001

Table 5. Results of pairwise Kruskal-Wallis tests

Clusters in comparison	*Chi-square*	*p-value*
Cluster 4 vs. Cluster 2	141.58	.000
Cluster 2 vs. Cluster 1	47.76	.000
Cluster 1 vs. Cluster 3	7.15	.007
Cluster 3 vs. Cluster 5	0.04	.842

We then conducted pairwise Kruskal-Wallis tests to determine if the difference in data quality between pairs of collaboration patterns was significant. As summarized in Table 5, the differences in quality between these patterns (except between cluster 3 and cluster 5) were statistically significant (p <0.01).

7. Discussion and Conclusion

Lih (2004) proposed two of the most widely used quality indicators of Wikipedia articles, "rigor" (number of edits) and "diversity" (number of unique editors). Although the idea that edits correspond to an increase in article quality is in general true, different contributors may contribute in different ways. Hence, the path to quality improvement may differ from one article to another. We believe that in essence, data quality depends on different types of contributors, i.e., the roles they play, and the way they collaborate.

In our research, we identified various roles a contributor may assume for a given article in the Wikipedia. Our research differs from (Anthony et al. 2005; Bryant et al. 2005) on two aspects. First, we define a contributor's role specifically for a given article, rather than for the Wikipedia community as a whole, since a role can vary from article to article. Secondly, we propose a novel approach to identifying roles by mining the

provenance, i.e., various actions carried out by a contributor on an article. Our research is also the first of its kind to identify collaboration patterns based on provenance in terms of *who* does *what.* We illustrate the impact of different collaboration patterns on data quality and identify patterns that are preferable or detrimental for quality: Articles developed using patterns where all-round editors played a dominant role are often of high quality, while patterns where starters and casual contributors dominate are often associated with low data quality.

Why do different collaboration patterns impact data quality differently? It is worth further studying the characteristics of different patterns and their impact on data quality. As a first step, we made two observations. First, a conspicuous problem with certain patterns is the lack of references in the articles. The reference ratio (ratio between the number of reference creations and sentence creations) is only 0.11 for articles in cluster 5 (where starters dominated sentence creations) while it is 0.42 for those in cluster 4 (where all-round editors dominated). A possible reason can be that the starters who dominated sentence creations for articles in cluster 5 tended to create sentences without citing sources, while other people often did not bother (or were unable), to identify the sources of these sentences. Secondly, articles developed in patterns where all-round editors dominated (including cluster 2 and cluster 4) have a much higher edit ratio (ratio between the number of sentence modifications and sentence creations) than those developed in other patterns. This is probably because unlike starters and content justifiers, all-round editors not only create sentences and justify them with links and references, but also modify the sentences created by themselves. This kind of "self policing" accounts for 31% of modifications made by all-round editors. The present quality control mechanism in the Wikipedia focuses on peer review. Our observations show that "self justifications" and "self policing" are equally important since it takes extra effort to add references and correct errors in sentences created by other people. Our observations call for a software tool that alerts contributors to justify their insertions by adding links and references. It is also necessary to develop mechanisms that motivate the contributors to revisit and modify their inserted sentences.

Our research goes beyond past research that attempts to automatically assess the quality of Wikipedia articles. Rather, our research focuses on understanding the relationship between collaboration patterns and data quality, thus providing insights to develop mechanisms that may guarantee and improve quality. We believe our research paves the way for developing new software tools for collaboration for the Wikipedia to encourage specific role setting and collaboration patterns to improve the quality of articles.

References

Anthony, D., S. Smith, T. Williamson (2005). "Explaining quality in Internet collective goods: Zealots and good Samaritans in the case of Wikipedia", Fall 2005 Innovation & Entrepreneurship Seminar, MIT.

Bryant, S. L., A. Forte, A. Bruckman (2005). "Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia", 2005 international ACM SIGGROUP conference on supporting group work: 1-10.

Denning, P., J. Horning, D. Parnas, L. Weinstein (2005). "Wikipedia risks", CACM 48(12): 152 - 152.

Giles, J. (2005). "Internet encyclopedias go head to head", Nature 438(7070): 900-901.

Lih, A. (2004). "Wikipedia as Participatory Journalism: Reliable Sources? Metrics for Evaluating Collaborative Media as a News Resource", 5th International Symposium on Online Journalism:16-17.

Liu, H. and V. Keselj (2007). "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests", Data & Knowledge Engineering 61: 304-330.

McGrath, J. (1984). Groups: Interaction and Performance. Englewood Cliffs, NJ, Prentice-Hall, Inc.

Niu, K., S. Zhang, J. Chen (2006). "An Initializing Cluster Centers Algorithm Based on Pointer Ring", Sixth International Conference on Intelligent Systems Design and Applications: 655 - 660.

Ram S., J. Liu (2007). "Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling", Lecture Notes on Computer Sciences 4512. Springer-Verlag: 17-29.