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.

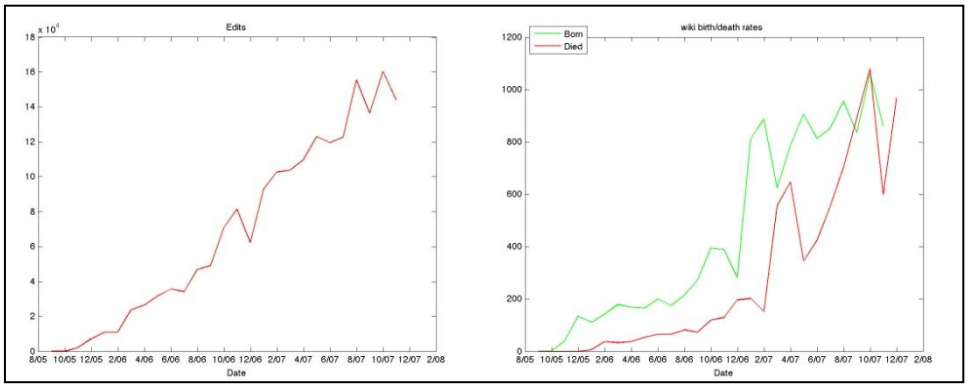
(f) **Visalization**. We generated two visualizations. First, we produced a density plot for each cluster, where we transformed wikis time series to an accumulator array, using a vector-to-raster conversion. By accumulatin*g* the number of lines passing through eac<u>h</u> array cell, we computed the 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

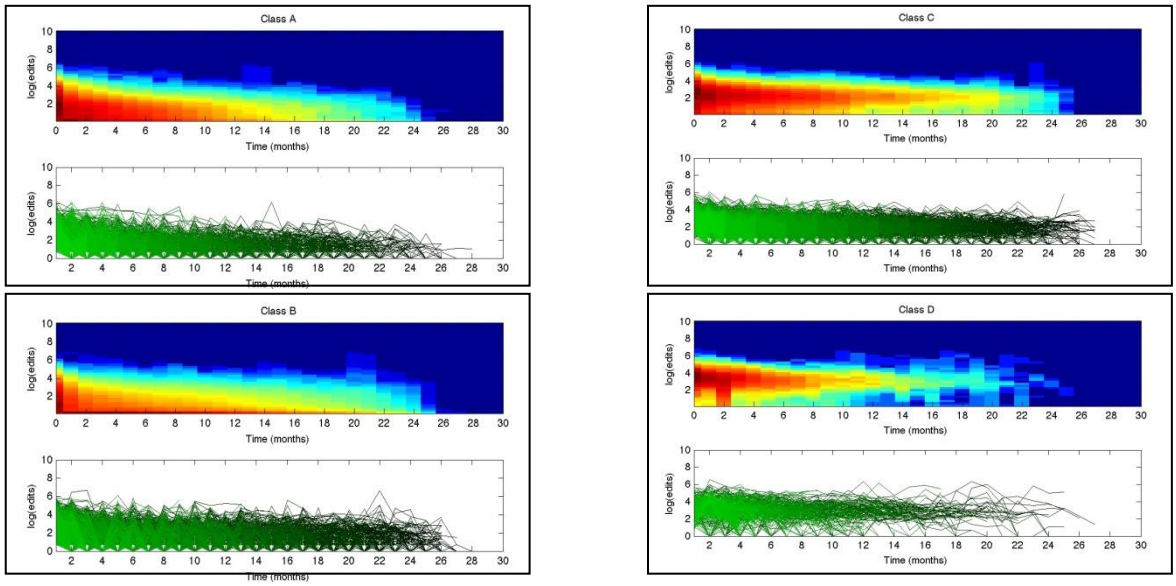
inactive shortly after their inception. We also notice that while at the initial period when wikis were introduced birth rates were rose continuously, towards the end of our analysis period TOI rates surpass 'birth' rates.



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.