A Utility for Estimating the Relative Contributions of Wiki Authors

Ofer Arazy[1], Eleni Stroulia[2]

School of Business[1], Dept. of Computing Science[2]
University of Alberta
Edmonton, AB T6G 2E8 Canada
oarazy@bus.ualberta.ca, stroulia@cs.ualberta.ca

Abstract

Wikis were originally designed to hide the association between a wiki page and the authors who have produced it. However, there is evidence suggesting that corporate wiki users require an attribution mechanism that would automatically record (and present) the relative contribution of each author. In this paper we introduce an algorithm for assessing the contributions of wiki authors that is based on the notion of sentence ownership. The results of an empirical evaluation comparing the algorithm's output to manual evaluations reveal the type of contributions captured by our algorithm. Implications for research and practice are discussed.

Introduction

Wiki, derived from the Hawaiian word for *fast*, is a web-based collaborative authoring application (Leuf & Cunningham 2001). In wikis, users can edit any part of the content of a wiki page. As a result, at any point in time, the most recent page version reflects the cumulative contributions of all users that have edited the page until then.

Wikis are originally designed to hide the association between a wiki page and the authors who have produced it (Leuf & Cunningham 2001). The main advantages of this feature are: (a) it eliminates the social biases associated with group deliberation, thus contributing to the diversity of opinions and to the collective intelligence of the group, and (b) it directs authors towards group goals, rather than individual benefits. However, this non-attribution is less suitable when users are motivated primarily by career-advancement goals. The main limitation of non-attribution is that it hampers accountability and reduces the motivations of wiki users to contribute content. As wikis are penetrating into corporate settings (Arazy et al. 2009), there is a need for an attribution mechanism that would automatically record (and present) the relative contribution of each author (Rashid et al. 2006).

In this paper, we discuss our initial work towards addressing this concern, and introduce a wiki add-on that automatically calculates the *relative contributions of wiki authors*. We expect that if an estimate of author contributions was presented on the wiki, corporate wiki users would be encouraged to participate in the collaborative authoring process.

Related Work

Several recent studies have proposed extensions to wikis to automatically calculate users' contributions and attribute a wiki page to its contributing authors. Hess et al. (2006) propose a utility that calculates the extent of a user's revision by comparing the current version to the previous one, such that the overall contribution of a user is based on the sum of all her revisions. Sabel (2007) uses a similar approach and proposes that the difference in revisions be used as a 'rating' of the author's contribution, which then feeds into a reputation system. Korfiatis et al. (2006) estimate a user's authority in the wiki based on social network analysis, propose that authors' centrality could serve as a proxy for their reputation. Ding et al. (2007) developed a utility to visualize wiki activity, where an author's contribution is based on a simple count of the edits he has made to wiki pages. Hoisl et al. (2007) implemented an add-on that measures the relative contribution of authors based on differences between revisions, weighting revision by their importance.

These recent works provide some interesting solutions to the problem of wiki attribution. However, they suffer from several limitations. First, social network approaches (Korfiatis et al. 2006) are good at estimating the distribution of one's efforts across wiki pages, but do not provide an estimation of the extent of contributions to a specific wiki page. Second, some of the proposed algorithms can easily be manipulated by users seeking to boost their contribution score, for example the ones using a simple count of page edits (Ding et al. 2007). Third, existing methods do not distinguish between contributions that remain on the wiki page over time and those that are quickly deleted (Hess et al. 2006; Sabel 2007). For example, if an author adds content that is deleted immediately, he still receives the same credit as if the contribution stayed on the wiki for months. Assuming that higher quality contributions are likely to persist, attribution algorithms should consider the duration a contribution stays on the wiki page. Lastly, rarely are the newly proposed algorithms evaluated to verify that they indeed

capture what they pertain to. Some of the concerns relate to the type of contributions that are captured (e.g. one algorithm may capture only new content contributions, while another may capture formatting changes to the page), and the extent to which the total contribution score of an author correctly represents the author's overall contribution to the wiki page.

The nature of contributions made to Wikipedia pages has been explored in several recent studies. Pfeil et al. (2008) tried to fill this gap by using a grounded theory to elicit a categorization of Wikipedia contribution types, which was later adopted by Ehmann et al. (2008). This categorization suggests that authors can make a contribution beyond adding new content, e.g. in formatting existing information or even by deleting irrelevant information. Their extensive categorization scheme is be too detailed for our purpose, as we are interested in identifying the nature of contributions captured by automatic algorithms and such algorithms may not be able to make the fine distinctions proposed in the framework proposed by Pfeil et al. (2008) and Ehmann et al. (2008). We build on this framework and propose a simpler categorization of the following contribution types: (I) *Add content*: adding complete new sections or changing existing information, (II) *Formatting*: changes affecting the structure or appearance of the page, (III) *Linking internally*: connecting to other pages on the same wiki, (IV) *Linking externally*: connecting to web pages outside the wiki, (V) *Delete*: deleting content, and (VI) *Proofread*: making minor corrections and refinements to text and hyperlinks. We are not aware of any studies that evaluated the association between attribution algorithm and the authoring categories.

Our Proposed Sentence-Ownership Algorithm

In our work, we set out to address the limitations discussed above by developing a wiki attribution algorithm that: (a) calculates the authors' contributions to each wiki page, (b) cannot be easily manipulated, (c) estimates the extent of a contribution using a sentence as the basic unit of meaning, and (d) distinguishes between contributions that persist on the page from those that are deleted. Our study assesses the algorithm by comparing it to human assessors' perceptions of contribution. The study investigates the category of contributions that the algorithm captures, as well as the correlation between the top contributors extracted by the algorithm and those identified by the human assessors.

The algorithm is based on the notion of *sentence ownership*, such that an author owns a sentence that he has created. Sentence-based metrics are used for calculating the number of sentences a user has added and the number of sentences he deleted. In addition, the algorithm calculates the number of internal and external hyperlinks created by the user, as well as the word-level changes made by each contributor. The most interesting innovation of the proposed algorithm is its sentence-ownership calculation component, as explained below.

We view a series of edits made by the same author as one continuous editing effort, and define a wiki page "release" as the last of these sequential revisions. The proposed algorithm calculates the sentence ownership of wiki page authors for each release. Our sentence-segmentation process translates the wiki markup into plain text by

(a) replacing links with their corresponding text,

(b) removing wiki formatting and macros,

(c) stripping out bullets and numbers from listings, and

(d) placing a full stop after individual list items.

Next, the algorithm invokes the sentence-segmentation tool of the UIUC Cognitive Computation Group[1] to rewrite the wiki page into a new page with one sentence per line.

Once sentence boundaries have been established, we compare sentences between the current and previous release, and consider a match when a sentence in the current release is very similar to one in the previous release[2]. It is possible that we identify multiple matches for a sentence in the current release, such that we face a many-to-many relationship between sentences in the new release and those in the old release, and we reduce those to one-to-one relations, such that each sentence in the current release has only one corresponding sentence in the previous release. Our algorithm uses the Munkres (1957) method to minimize the total distance between the paired sentences, where in essence the distance is an estimate of how much the sentence's position has changed relative to its established context. If a sentence in the current release does not have a match in the earlier release, it is considered as newly added sentence with the author as the sentence owner. If, conversely, a matching sentence is found in the previous release, we distinguish between a minor change[3] and a major change[4]. For minor changes, if after the changes the original owner is still responsible for creating more than 50% of the words in the sentence, then he still retains ownership of that sentence. On the other hand, if after the changes the original owner is now responsible for creating less than 50% of the words in the new sentence, the original owner looses ownership and the sentence becomes public. For major changes, the contributor responsible for these changes becomes the new owner of the sentence. We explore two variations of the algorithms. The first calculates the total amount of work (e.g. sentences owned, hyperlinks created) a contributor has made. The second considers only the contributions that persist in the most recent version, assuming that contributions that persist are more relevant. We calculate the two variations for all metrics, but for deleted sentences.

---

[1] http://l2r.cs.uiuc.edu/~cogcomp/atool.php?tkey=SS

[2] We consider a "match" if more than 50% of the words of $S_{previous}$ can also be found in $S_{current}$.

[3] We consider a "minor change" if *more than* 50% of the words of $S_{current}$ can also be found in $S_{previous}$.

[4] We consider a "major change" if *less than 50%* of the words of $S_{current}$ can also be found in $S_{previous}$.

Empirical Evaluation of Algorithms

In order to assess the accuracy in which the proposed algorithm captures contributions, we compared it against manual assessments. We analyzed nine random Wikipedia articles. The average article was included over 130 distinct contributions, made by 105 unique authors. Two research assistants analyzed each edit made to each article independently, by reviewing the 'History' section of articles and comparing subsequent versions. For each article, the assessors performed two types of manual assessments. First, the assessors examined each of the edits made to the article by recording the author and classifying the contribution type into the categories mentioned in Section 2. The extent of each contribution was then ranked on a 1-5 scale, from Minor to Major. After completing the analysis of a page, we've summed up the contributions by author, giving each author an overall page score for each of the contribution categories. Second, the assessors have identified the top contributors (up to 5) and ranked them in order of their contribution. The analyses were performed independently by the two assessors, and we used the average of the two assessors in our evaluation.

In order to investigate the extent to which the various automatically calculated metrics are associated with the collaborative authoring categories, we performed a Pearson correlation analysis. The results are described in Table 1 below. For the purpose of this analysis we compared the algorithm's results for the *total amount of contributions* made by the author. We found that the manual identification of top page contributors was most highly correlated with the 'Add' class (correlation = 0.65). Other types of contributions that seem to impact the perception of a top contributor are 'Link' (correlation = 0.24) and 'Structure' (correlation = 0.15). 'Delete' and 'Proofread' edits had little impact on perceptions of top contributors.

Our analysis shows that there is a strong association between algorithms and specific authoring categories they intend to capture. The number of sentences owned is mostly correlated with the 'Add' category; the link count is associated strongly with the authoring categories of internal and external links; and the count of deleted sentences is highly correlated with assessors' perceptions of the extent of deletions. These associations are marked with a gray background.

However, some automatic metrics are strongly correlated with more that one of the authoring categories. For example, the number of sentences owned is highly correlated with the 'Format' and 'Internal Link' categories; internal link count is strongly correlated with the 'Add' and 'Format' categories. We suspect that this is due to the fact that authoring categories are correlated amongst themselves, such that an active contributor makes many contributions along several categories. The table illustrates that if one is interested in capturing the full extent of authors' contributions, no one metric is sufficient, e.g. sentence ownership is effective at capturing 'Add' and 'Format', but it is not useful at capturing 'Proofread' and 'Delete' contributions. Thus, a combination of several metrics is required.

Table 1: correlations of manually identified edit categories with the various automatically calculated metrics. All correlations are statistically significant at P<0.001 (using a 2-tailed t-test), except where indicated by '*'.

	Manual Analysis					
Automatic Algorithm	Add	*Format*	Int. Link	Ext. Link	Proofread	Delete
Sentences Owned	*0.39*	*0.35*	*0.29*	0.13	0.18	0.08
Internal Links Count	**0.48**	**0.55**	**0.64**	0.14	*0.28*	0.19
External Links Count	*0.24*	*0.32*	0.15	0.52	0.12	0.09
Word-level Changes	0.10	*0.35*	**0.53**	0.12	**0.46**	*0.27*
Deleted Sentences	0.15	*0.24*	0.19	0.05*	0.15	*0.34*

When comparing the automatic metrics to the top page contributors as perceived by our assessors, we find – surprisingly – that the metric that is correlated with the assessors' perceptions of top contributors is the internal link count. The baseline (a simple count of edits), too, is unexpectedly highly correlated with assessors' perceptions. The results are presented in Table 2 below. It is important to note that, although some metrics (edit count, link count) seem to be effective at identifying the overall top contributors, these metrics could easily be manipulated by users trying to promote themselves. Our sentence ownership algorithm still performs well and is less sensitive to such manipulation.

Table 2: correlations of the various metrics with assessors' perceptions of top contributors. All correlations are statistically significant at P<0.001 (using a 2-tailed t-test).

Baseline: # of Edits	Sent. Owned	Int. Link Count	Ext. Link Count	Word-level Changes	Deleted Sent.
0.48	0.31	0.54	0.32	0.24	0.20

If one was interested in using a combination of metrics for identifying the top contributors, a stepwise regression shown that the optimal combination includes three metrics: internal link count, word-level changes, and sentences owned. These three variables together are able to explain 30% of the variance of the dependent variable, 'Top Contributor' (adjusted $R^2 = 0.300$; F = 135; t = 0.000). The coefficients of the three variables are statistically significant (t values are 14.8, 3.4, and 2.6 respectively). It is interesting to note that the number of edits is not included in the list of variables, probably because it is redundant, as the three variables that are included capture the majority of edits.

Our final analysis concerns the comparison of metrics that count a user's overall contributions against metrics that capture only those contributions that remain at the current version. Table 3 shows how these two types of metrics are correlated with assessor's perceptions of top contributor. Overall, we see no big difference between the two approaches. This was unforeseen, as we expected the metrics for the current version to indirectly capture the quality of contributions and thus to be more correlated with perceptions of top contributors. We find that the 'current version' metrics perform better for sentence ownership and the internal link count, while 'total contributions' metrics are better for the external link count and word-level changes.

Table 3: comparing metrics for total contributions against metrics for the contributions that persist in the most recent version. Both are correlated with assessors' perceptions of top contributors. All correlations are statistically significant at $P<0.001$ (using a 2-tailed t-test).

Correlations with Top Contributor Perceptions	Sent. Owned	Int. Link Count	Ext. Link Count	Word-level Changes
Current version	0.33	0.26	0.51	0.30
Total contributions	0.31	0.24	0.54	0.32

Discussion and Conclusion

While non-attribution is useful in promoting democratic deliberation on the internet, it prevents corporate users from gaining recognition for their wiki work. Recent studies have proposed software utilities that would automatically attribute a wiki user with a score representing his contribution. However, these proposed algorithms suffer from several drawbacks, as they often use course measures, they are easy to manipulate, and they often capture just a sub-set of the classes of contributions.

In this paper we've tried to address these gaps by proposing a novel wiki attribution algorithm and comparing it against human perceptions. The innovation of our algorithm lies in (a) the sentence-ownership algorithm, and (b) in calculating contributions that persist in the current version of the wiki page (in addition to metrics calculating the overall contribution). We argued that the count of sentences that survived the wiki process of continuous refinements implicitly captures the quality of a user's contributions.

One of the most surprising result of our study is that the metric that is most correlated with assessors' perceptions of top contributors is the internal link count. We do not believe that assessors' perceptions were strongly affected by the number of internal links an author makes. Rather, we explain this result by the fact that the ones adding links are active across a variety of categories, and this is why they are perceived as top contributors. We were also surprised to find that the simple edit count – used as a baseline – performed very well, yielding higher correlation with top contributor perceptions that other metrics such as sentence ownership. We believe that this is due to the fact that the edit count captures range of contributions across all categories, while the other metrics are associated with only a sub-set of the authoring categories. The sentence ownership metric performed fairly well, and has the advantage that it is less vulnerable to manipulations.

Additional research is warranted in order to explore the design of more advanced wiki attribution algorithms, so that we can gain a better understanding of the authoring categories captured by various algorithms, and assess whether the presentation of user attribution indeed motivates wiki users to enhance their participation.

References

Arazy O., Gellatly I., Jang S., and Patterson R. (2009) Wiki Deployment in Corporate Settings: A Case Study, *IEEE Technology and Society*, forthcoming.

Ding, X., Danis, C., Erickson, T., & Kellogg, W. A. (2007). Visualizing an enterprise wiki. In *Proceedings of ACM CHI '07,* 2189-2194, San Jose.

Ehmann K., Large A., and Beheshti J., 2008, Collaboration in context: Comparing article evolution among subject disciplines in Wikipedia, *First Monday*, 13:10, 6 October 2008.

Hess, M., Kerr, B., & Rickards, L. (2006). *Wiki user statistics for regulating behaviour*, working paper.

Hoisl, B., Aigner, W., & Miksch, S. (2006). Social rewarding in wiki systems - motivating the community. *Lecture Notes in Computer Science , 4564*, 362-371.

Korfiatis, N. T., Poulos, M., & Bokos, G. (2006). Evaluating authoritative sources in collaborative editing environments. *Online Information Review, 30* (3), 252-262.

Leuf, B., & Cunningham, W. (2001). *The Wiki Way: quick collaboration on the web.* Addison-Wesley.

Munkres, J., 1957, Algorithms for the assignment and transportation Problems. *Journal of the Society for Industrial and Applied Mathematics (SIAM)*, 32-38.

Pfeil U., Zaphiris P., and Ang C.S., 2008, Cultural Differences in Collaborative Authoring of Wikipedia**,** *J. of Computer-Mediated Communication*, 12:1, pp. 88-113.

Rashid A., Ling K., Tassone R., Resnick P., Kraut R. and Riedl J. (2006). Motivating Participation by Displaying the Value of Contribution. *Conf. on Human Factors in Computing Systems*. NY, USA. April 22-27, 2006. 955-958

Sabel, M. (2007). Structuring wiki revision history. *Proceedings of the 2007 international symposium on wikis* (pp. 125-130). Montréal: ACM.