A Comparison of Generated Wikipedia Profiles Using Social Labeling and Automatic Keyword Extraction

Terrell Russell[1], Bongwon Suh[2], Ed H. Chi[2]

[1]University of North Carolina at Chapel Hill, Chapel Hill, NC
[2]Palo Alto Research Center, Palo Alto, CA

Abstract

In many collaborative systems, researchers are interested in creating representative user profiles. In this paper, we are particularly interested in using social labeling and automatic keyword extraction techniques for generating user profiles. Social labeling is a process in which users manually tag other users with keywords. Automatic keyword extraction is a technique that selects the most salient words to represent a user's contribution. We apply each of these two profile generation methods to highly active Wikipedia editors and their contributions, and compare the results. We found that profiles generated through social labeling matches the profiles generated via automatic keyword extraction, and vice versa. The results suggest that user profiles generated from one method can be used as a seed or bootstrapping proxy for the other method.

Introduction

Given the rise of Web2.0 and user-generated content, there is a great need to understand how to represent and summarize a user's actions on a website. Understanding and categorizing patterns of user contributions can be particularly useful because they can be used to help judge the interests and potentially the areas of expertise of the user.

One way to summarize users' contributions is to create keyword profiles that describe the contributions that they have made. In this paper, we examine two different profile generation techniques and apply these techniques on the contributions made by active Wikipedia users.

An emerging approach to generating summary profiles is through users describing other users with particular keywords or tags (Farrell et al. 2007; Razavi and Iverson 2008) – a process that we call "social labeling". The top half of Figure 1 shows an example of a tag cloud generated by collecting social labels for an active editor in Wikipedia. The keyword profiles obtained through social labeling can be thought of as a summary representation of the user's edits and interests (Alonso, Devanbu and Gertz 2008; Razavi and Iverson 2008). This approach is simple, clean, and scalable, however, the main disadvantage is that this method requires human labor and is therefore fairly expensive in time and/or money.

Automatic keyword extraction techniques, on the other hand, use word count statistics to select the most salient keywords from the contributions to represent the user profile. The bottom half of Figure 1 shows an example of a tag cloud generated using this method. The advantage is that this technique does not involve manual effort; however, it is not clear that this method generates sufficiently good summarizations.

Figure 1: Two tag clouds representing the same Wikipedia editor. The upper cloud consists of social labels collected from readers, while the bottom cloud consists of keywords that were automatically collected from the revisions made by the editor.

In this study, we compare the results of social labeling processes to an automatic keyword extraction technique for building profiles that summarize active Wikipedia user interests and/or areas of expertise. We present the results of this study and conclude with some design implications.

Two Methods For Generating a User Profile

We generated two types of user profiles for highly active Wikipedia editors by (1) collecting and aggregating the labels generated by human readers and (2) selecting salient keywords that were automatically extracted from revisions that the editor contributed.

We chose four articles from the English Wikipedia (Evolution, Coffee, NFL, and Shotgun) and for each article we chose three editors who made the most number of edits on the page giving us twelve total editors of interest. We excluded editors who are Wikipedia administrators, have no user page on the Wikipedia site, or have made less than 500 total Wikipedia edits as the human evaluators would need enough data to make informed judgments and the automatic extraction algorithm would have enough data to generate lists of sufficient size. Administrators were excluded as their interests are usually Wikipedia itself and potentially conflate the analysis. We built two user profiles for each of the twelve editors by applying the following two methods.

Method 1: Social Labeling

In order to obtain humans' perceived judgment on Wikipedia editors' behavior, we invited users of the Mechanical Turk system, or *turkers*, to read selected pages about Wikipedia editors. Amazon's Mechanical Turk (http://www.mturk.com) is a market in which simple tasks are posted and small monetary rewards are paid for completing them. We asked the *turkers* to complete a short survey about their findings concerning the Wikipedia editor and their contributions.

Wikipedia Page About Editors	Description
1. User page	Functions like a home page for an editor and primarily edited by the editor
2. User Talk page	A discussion page where other Wikipedia editors communicate with this editor (Figure 2b)
3. Contributions	Lists all revisions made by the editor in reverse chronological order
4. User page + WikiDashboard	User page with WikiDashboard embedded at the top (Figure 2a)
5. User Talk page + WikiDashboard	User Talk page with WikiDashboard embedded at the top

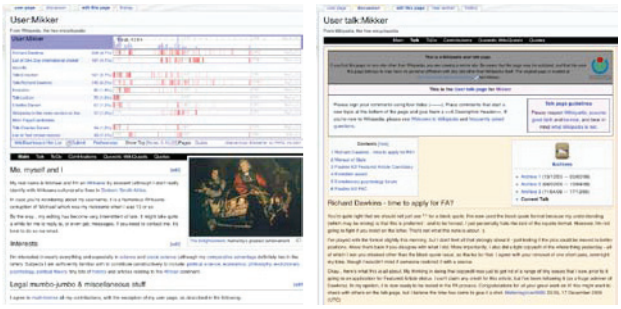
Table 1: Five types of Wikipedia pages used for social labeling. The first three types are directly provided by Wikipedia. The last two are augmented versions of the first two.

Wikipedia holds encyclopedic entries called article pages. In addition to its article pages, Wikipedia also provides supplementary types of pages to assist community functions such as article discussion, categorization, and general discussions around policy, norms, and decorum (http://en.wikipedia.org/wiki/Wikipedia:Namespace).

Among them, *user pages* and *user talk pages* present information about editors and facilitate communication between them. As the need for coordination increases, these non-article pages have been serving a critical role in building communities in Wikipedia (Kittur et al. 2007).

In this study, five different types of Wikipedia pages were provided to *turkers* to help them label the editors (Table 1). We used three basic types of pages supported by Wikipedia natively as well as two additional types of pages (the last two types in Table 1). Figure 2 shows two examples of what was shown to *turkers*.

In the two additional types, a dashboard visualization was embedded into pages (e.g. Figure 2a) to present additional social dynamics and the editing patterns of an editor (Suh et al. 2008). These two types were included to examine whether the dashboards affect user performance, which is part of a separate on-going study.



(a) User page + WikiDashboard (b) User Talk page

Figure 2: Two example pages used in the study. The study participants were asked to label the editor whose activities are described in the pages.

For each task, we asked the study participants to take a couple minutes to familiarize themselves with the content on the page and requested they type five labels (or keywords or tags) that represent the Wikipedia editor (e.g. bird, apple). As discussed in Kittur (Kittur et al. 2007), when dealing with anonymous Mechanical Turk users, it is very important to have explicitly verifiable questions as part of the task. We required *turkers* to answer a number of questions to convince us they were human and to force them to process the contents of what they were being shown. For example, when a participant was provided with a User Talk page (Figure 2b), we asked them to type the name of the user who made the most recent edit to ensure that they were cognitively aware of what they were seeing.

As mentioned earlier, twelve Wikipedia editors were chosen for this study and five types of user pages were prepared for each editor. We assigned ten *turkers* to each of the 60 (12x5) conditions for a total of 600 task responses. Study participants were advised to finish just one task to increase the diversity in the sample of *turkers*. We also collected some demographic information about their own Wikipedia experience in a simple survey form. The participants were paid $0.10 per response.

Method 2: Keyword Extraction

As designed, wikis archive entire edit histories so that any individual revision can be retrieved later. We built profiles for the editors of interest by analyzing archived revisions of pages they had edited, extracting their edits, and then synthesizing the results.

Since each revision is saved in Wikipedia, it is possible to identify (1) who is responsible for each revision and (2) exactly which part of a document was added, removed, and/or relocated for that revision. We retrieved all edits made by the selected editors from a dump file of English Wikipedia containing a total of 167.4 million revisions. We utilized the Hadoop distributed computing environment (http:// hadoop.apache.org) to store, parse, and analyze the data.

We excluded common English and Wikipedia specific stop words (e.g. "user", "page", html tags, the names of months) from analysis. Since we aimed to collect keyword tags that represent the editors, we chose to disregard words deleted or relocated by editors and only used words added by those editors. However, further research is required to investigate the validity of this choice. We collected a set of tokenized words that were added by the editors, resulting in [editor, word, frequency] tuples. The amassed tuples can be transformed to a tag cloud as shown in Figure 1.

Results

We examined the similarity between the user-generated social labeling profile (perceived profile) and extracted keywords profile (behavioral profile) for each of our twelve Wikipedia editors.

To generate our social labeling profiles, 313 *turkers* participated in the study. They completed 526 tasks and we were able to collect 2,521 total tags (848 unique tags). Even though we discouraged multiple participations from any single *turker*, 52 *turkers* finished more than one task. Among these 52 *turkers*, we found three who blatantly vandalized the survey and their results are excluded (425 tags, 17% of the total tags) for the analysis. All together, we obtained 1,138 [editor, word, frequency] tuples.

Regarding the automatic keyword extraction profile, we collected 116,430 [editor, word, frequency] tuples.

Social Labeling vs. Keyword Extraction

To investigate the correlation between the two sets of editor profiles, we performed Spearman's rank correlation analysis for each pair of the profiles. The analysis generated a set of correlation measures (Spearman's *rho*) and p-values on their statistical significance (Table 2). Approximately, Spearman's *rho* tells us whether the two methods rank keywords in roughly the same order. The profiles generated by the *turkers* were significantly smaller than those extracted automatically. The rank correlation was done on the intersection of each pair of profiles as cursory inspection suggested the vast majority of the *turker* labels were present in the automatically extracted keywords. The average length of the intersected ranked lists was 59 words.

As shown in Table 2, among twelve editors, ten profile comparisons (along the diagonal) show a significant correlation ($p < 0.05$). Given sufficient data input sizes, this analysis shows that the *turkers* were able to tag the editors with words that seem to reflect their editing word choices by simply looking at the Wikipedia user pages. And vice versa, as this is a correlation, the keyword extraction algorithm seems to be able to construct reasonable profiles that reflect how others might perceive the editors' interests.

	Social Labeling Profile											
Keyword Extraction Profile		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
	K1	**.55**	**.27**	**.37**	.41	.08	.21	.13	.49	.13	.06	.27	.24
	K2	.25	**.29**	**.32**	.01	.21	.13	.29	.11	.14	.33	.14	.17
	K3	.14	**.26**	**.33**	.24	.05	.03	.47	.32	.07	.03	.30	.24
	K4	.12	.02	.21	**.28**	.10	.17	.16	.14	.26	.03	.25	.09
	K5	.26	.12	**.42**	.01	.29	.04	.09	.20	.24	**.44**	.24	**.39**
	K6	.04	.08	.03	.14	.21	**.40**	.23	.43	.05	.02	**.31**	.20
	K7	.18	.24	.08	.28	.40	.14	**.44**	.22	.14	.14	.07	.04
	K8	.18	.07	.15	.18	.32	.01	.21	**.51**	.29	.11	.12	.16
	K9	.03	.10	.17	.21	.28	.01	.08	.08	**.48**	.00	.27	.33
	K10	.01	.06	.18	.13	.13	.15	.01	.46	.16	.20	.11	**.30**
	K11	.16	.15	.25	.08	.10	.07	.11	.18	.06	.10	**.27**	**.49**
	K12	.09	.09	.20	.26	.18	.03	.26	.40	.07	.00	.10	**.62**

Table 2: Spearman's rank correlation (rho) for the twelve Wikipedia editors. Shaded cells indicate significant correlation ($p<0.05$) between a social labeling profile (S1 S12) and a keyword extraction profile (K1 K12). 10 of 12 corresponding pairs of profiles were significant (along the diagonal).

While the analysis suggests strong compatibility between the two profiling methods, we also would like to confirm that the social labeling also generates a reasonably unique profile to each editor. We constructed the similarity matrix between all profiles (12 social labeling profiles X 12 keyword extraction profiles) seen in Table 2.

The result shows that the profiles are indeed reasonably unique, as shown in the corresponding shading pattern. Greater distinguishing performance should be attainable through further research. Overall, these analyses show very strong compatibility between the social labeling and keyword extraction techniques.

Discussion and Limitations

The study results suggest that automatically generated behavioral user profiles can be a reasonable approximation to a human-judged perceived user profile. This suggests that the automatic method generates good summaries of users' activities for sites with user-generated content, given enough user generated content is available for analysis. The results suggest that the keyword extraction method can

be a useful bootstrap when one builds a social labeling system and wants to pre-populate the user profiles. Given the high cost of manual input labor, this bootstrapping could be a time and effort saver for systems such as enterprise document repositories.

Interestingly, during this study, we found many cases where the Mechanical Turk users complained that they did not have enough information to judge editors' areas of knowledge or expertise. However, the aggregation of their individual inputs was sufficient to successfully build a reasonably complete profile, which is a typical exemplar of how collaboration systems work.

User profiles are also useful for expertise location. Some research shows that users' activity profile could be useful for judging expertise (Alonso, Devanbu and Gertz 2008; Farrell et al. 2007; Razavi and Iverson 2008). McDonald (McDonald and Ackerman 1998) describes a field study, using richer data than word lists, of how people find others with expertise and people's tendency to form agreement when judging each other's expertise. However, Wattenberg et al. (Wattenberg, Viegas, and Hollenbach 2007) show that Wikipedia administrators show differing edit patterns, categorizing their edits into either systematic or reactive activities. In our study, profiles are summaries of active user contributions but we do not investigate the nature of the edits beyond tokenization. Looking at these keyword profiles might enable readers to judge the interests and possibly the areas of expertise of these editors.

Among numerous research on online trust and reputation (Jøsang, Ismail, and Boyd 2007), perhaps the most similar work to this study is Bogers et al. (Bogers, Thoonen, and van den Bosch 2006). They investigated automatic methods of expertise extraction and evaluated it using a baseline of human experts' judgment. They showed that people are able to estimate expertise of each member of the workgroup from the aggregated content of his or her publications. In contrast, we applied a *diff*-based approach on the Wikipedia archive and constructed editor profiles from that editor's wiki contributions.

One of technical issues that needs to be addressed is handling multi-word tags. The keyword extraction method tokenized strings using white space as the delimiter. Thus, the keyword profile is composed of single words (e.g. "Richard" and "Dawkins"). However, when collecting tags from *turkers*, no such restriction was imposed and a reasonable number of inputs (540 out of 2096) were multi-word tags (e.g. "Richard Dawkins"). For the comparison done in this paper, multi-word tags were then broken into multiple single tags for consistency. Further research is needed to deal with bigram and trigram keywords.

Other important design issues of social labeling include privacy, vandalism, and a potential gaming of the system. Exposing information about a contributor in open participation systems could have a negative impact. An important question for future research is how to face these challenges.

Conclusion

In this paper we investigated two profiling techniques for summarizing user contributions – social labeling and automatic keyword extraction from a user's Wikipedia activity. The analysis moderately confirms that profiles generated through social labeling match the contribution patterns of active Wikipedia editors, suggesting compatibility between the two profile generation methods. The results suggest that social labeling has a strong potential for impacting online collaboration systems by improving the identity and awareness among users, and therefore, the quality of user-generated content. We hope this finding may inform designers seeking to construct up-to-date profiles for users of collaborative systems.

References

Alonso, O., Devanbu, P.T., and Michael Gertz, M. Expertise identfication and visualization from CVS, In *Proc. MSR '08* (2008), 125-128.

Bogers, T., W. Thoonen, W., and van den Bosch, A. Expertise Classification: Collaborative Classification vs. Automatic Extraction, In *Proc 17th SIG/CR Classification Research Workshop*, (2006).

Farrell, S., Lau, T., Nusser, S., Wilcox, E., and Muller, M. Socially augmenting employee profiles with people-tagging, In *Proc. UIST '07* (2007), 91-100.

Jøsang, A., Ismail, R., and Boyd, C. A Survey of Trust and Reputation Systems for Online Service Provision, *Decision Support Systems* 43 (2007), 618-644.

Kittur, A., Suh, B., Pendleton, B.A., and Chi., E.H. He Says, She Says: Conflict and Coordination in Wikipedia. In *Proc. CHI 2007*, ACM Press (2007), 453-462.

Kittur, A., Chi, E. H., Suh, B. Crowdsourcing user studies with Mechanical Turk. In *Proc. CHI 2008*, ACM Press (2008), 453-456.

McDonald, D.W. and Ackerman, M.S. Just Talk to Me: A Field Study of Expertise Location, In *Proc CSCW 98*, (1998), 315-324.

Razavi, M.N. and Iverson, L., Supporting selective information sharing with people-tagging, In *extended abstract CHI '08*, 3423-3428.

Suh, B., Chi, E.H, Kittur, A., Pendleton, B.A. Lifting the Veil: Improving Accountability and Social Transparency in Wikipedia with WikiDashboard. In *Proc. CHI 2008*, ACM Press (2008), 1037-1040.

Wattenberg, M., Viegas, F.B., and Hollenbach, K. Visualizing Activity on Wikipedia with Chromograms. In *Lecture Notes in Computer Science, Human Computer Interaction INTERACT 2007* (2007), 272-287.