A Lightly Supervised Approach to Role Identification in Wikipedia Talk Page Discussions

Oliver Ferschke
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
ferschke@cs.cmu.edu

Diyi Yang
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
diyiy@cs.cmu.edu

Carolyn P. Rosé
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
cprose@cs.cmu.edu

Abstract

In this paper we describe an application of a lightly supervised Role Identification Model (RIM) to the analysis of coordination in Wikipedia talk page discussions. Our goal is to understand the substance of important coordination roles that predict quality of the Wikipedia pages where the discussions take place. Using the model as a lens, we present an analysis of four important coordination roles identified using the model, including Workers, Critiquers, Encouragers, and Managers.

Introduction

Resources such as Wikipedia are successful as a result of the coordinated efforts of the masses. Despite its success, however, the quality and thoroughness of Wikipedia articles is variable depending upon the topic area and the popularity of the respective subject. Research into the reasons behind the variability point to issues of coordination. For example, Kittur and Kraut (2008) have shown in a cross-sectional analysis of Wikipedia articles and Talk pages that explicit coordination through discussion is particularly valuable in the early stages of an article's development. More broadly, high quality pages are associated with a history in which a small number of editors coordinated effectively with one another either directly (through discussion on the talk pages) or indirectly (through joint editing of textual artifacts). However, this research only focused on transactional data and did not analyze in detail what precisely was the substance of the coordination. Qualitative analysis of behavior in Wikipedia has begun to identify some important roles (Welser et al. 2011), although these identified roles are not exhaustive and have not been formally operationalized or validated through quantitative methods.

In this paper, we take the next step in operationalizing and validating these coordination roles through a lightly supervised role identification model (RIM) developed in previous work (Yang, Wen and Rosé, submitted). We apply this model to the dataset used in the earlier work of Kraut and Kittur (2008). In that study, the authors observed the development of articles with different quality ratings according to the WikiProject article quality grading scheme[1] in fixed time windows of 6 months and looked at the discussion activity during that time. In order to gain insights into the actual content of these discussions, we randomly sample 6,000 articles from the dataset and retrieved the discussions posted in each 6-month time window defined in the initial study from the Talk page revision history using the Wikipedia Revision Toolkit (Ferschke, Zesch, and Gurevych 2011). There is reason to believe that there are important roles for editors to play in a coordinated way in the trajectory of a page, and that detection of these roles and the configurations of roles in conversations, which could be seen as the composition of teams working towards article improvement, should predict page quality as the article evolves over time. The RIM model enables us to test this hypothesis through application to the extract from the Kraut and Kittur dataset.

In the remainder of the paper we review prior computational work on modeling role based behavior and briefly describe the RIM model that we recently introduced (Yang, Wen and Rosé, submitted). We then describe in more detail how we prepared the Wikipedia data for modeling. Next we describe the experiment that we ran and its results. We conclude with directions for our continued research.

Related Work

The concept of "social role" has long been used in social science fields to describe the intersection of behavioral, symbolic, and structural attributes that emerge regularly in particular contexts. Another similar line of work studies identification of personas (Bamman, O'Connor, and Smith 2013; Bamman, Underwood, and Smith 2014), e.g., celebrity, newbie, lurker, flamer, troll and ranter, etc. within the context of a social network, which evolve through user interaction (Forestier et al. 2012). What is similar between stances and personas on the one hand and roles on the other is that the unit of analysis is the person. On the other hand, they are distinct in that stances (e.g., liberal) and personas (e.g., lurker) are not typically defined in terms of what they are meant to accomplish, although they may be associated with kinds of things they do. Teamwork roles are defined in terms of what the role holder is meant to accomplish.

The notion of a natural outcome associated with a role suggests a modeling approach utilizing the outcome as light supervision towards identification of the latent roles. For example, Hu et al. (2009) predict the outcome of featured ar-

[1] http://en.wikipedia.org/wiki/WP:ASSESS

ticle nominations based on user activeness, discussion consensus and user co-review relations. The latent construct of a role, such as team leader, is defined in terms of a distribution of characteristics that describe how that role should ideally be carried out. Roles need not only be identified with the substance of the text uttered by role holders. Previous work discovers roles in social networks based on the network structure (Hu and Liu 2012; Zhao et al. 2013). However, these approaches do not standardly utilize an outcome as supervision to guide the clustering.

Role Identification Model (RIM)

Our role identification model aims to maximize the predicted quality scores of teamwork among a selected set of editors. We first introduce the basic notation and then present an iterative model that alternates between the following two stages: (i) teamwork quality prediction using article quality as an outcome variable, and (ii) participant role matching.

Notation

Suppose we have C teams in which participants collaborate to plan editorial work on an article. The number of participants in the j-th team is denoted as $N_j, (1 \leq j \leq N_j)$. There are K roles across C teams that we want to identify, where $1 \leq K \leq N_j, \forall j \in [1, C]$. That is, the number of roles is smaller than or equal to the number of participants in a team, which means that each role should have one participant assigned to it, but not every user needs to be assigned to a role. Each role is associated with a weight vector $W_k \in \mathcal{R}^D$ to be learned, $1 \leq k \leq K$ and D is the number of dimensions. Each participant i in a team j is associated with a behavior vector $B_{j,i} \in \mathcal{R}^D$. The measurement of teamwork quality is denoted as Q_j for team j, and $\hat{Q}_j$ is the predicted quality. Here, $\hat{Q}_j$ is determined by the inner product of the behavior vectors of participants who are assigned to different roles and the corresponding weight vectors.

Teamwork Role Identification

Our goal is to find a proper teamwork role assignment that positively contributes to the teamwork outcome (i.e. improvement of article quality) as much as possible. The role identification process is iterative and involves two stages. The first stage uses an regression model to adjusts the weight vectors in order to predict the teamwork quality, given a fixed role assignment that assumes participants are well matched to roles. In the second stage, we iterate over all possible assignments to find a matching that maximize our objective measure. In order to avoid the complexity of a brute force enumeration, we create a weighted bipartite graph and apply a maximum weighted matching algorithm (Ravindra, Magnanti, and Orlin 1993) to find the best matching. For each team (i.e. for each discussion topic), a separate graph is created as shown in Figure 1. We alternate between the two stages until both role assignment and teamwork quality prediction converge.

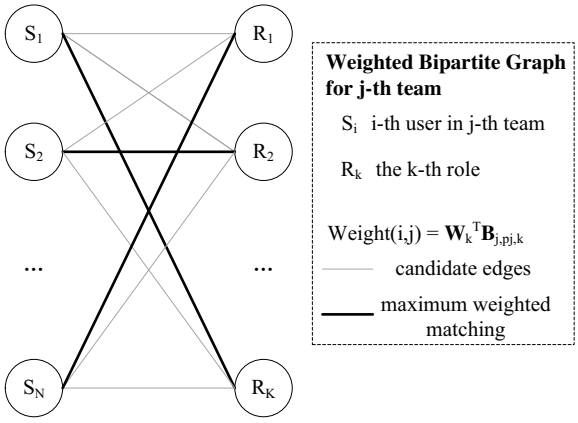

Figure 1: Weighted bipartite graph for matching users and roles

Assembling Behavior Vectors

We operationalize roles as distributions of behaviors that an editor engages in during participation in Talk Page Discussions. In order to represent Talk Pages in terms of the represented roles, we must first define types of contributions to such discussions. Then we must associate each participant in each discussion with a distribution of such behaviors based on their observed contributions. In this section we first describe a coding scheme that describes types of contributions to Talk Page discussions developed in earlier work Ferschke, Gurevych, and Chebotar (2012). Then we describe the technical process of segmenting Talk Page discussions, assigning contribution types to each segment, and constructing behavior vectors for each participant in each discussion.

Coding Scheme for Talk Page Contributions

In order to capture the discussion behavior of the individual users, we follow the approach described by (Ferschke, Gurevych, and Chebotar 2012), who developed a dialog act annotation scheme for Wikipedia discussions with labels that identify the intentions of discussion posts with respect to their relevance for the article improvement. We use a revised version of this annotation scheme (Ferschke 2014), which contains twelve dialog act labels in four categories shown in Table 1. *Article criticism* labels identify turns that mention article deficiencies such as missing information, factual or stylistic error, missing citations or layout problems. *Self commitment* labels indicate the willingness of the poster to improve the article or reports of them already having done so. *Requests* capture suggestions for edits by other community members. Finally, label in the *Interpersonal* capture the attitude of the posters towards each other. We train Naive Bayes classifiers on a manually labeled corpus of English Wikipedia discussions (Ferschke 2014) using the set of features described by Ferschke et al. (2012). We use these classifiers to label all turns in our dataset. The labels are described in detail by Ferschke (2014).

Label	Description
CRITCOMPL	Information is incomplete or lacks detail
CRITACC	Lack of accuracy, correctness or neutrality
CRITLANG	Deficiencies in language and style
CRITSUIT	Content not suitable for an encyclopedia
CRITSTRUCT	Deficiencies in structure or visual appearance
CRITAUTH	Lack of authority
ACTF	Commitment to action in the future
ACTP	Report of past action
REQEDIT	Request for article edit
REQMAINT	Request for admin or maintenance action
ATTPOS	Positive attitude
ATTNEG	Negative attitude

Table 1: Dialog act labels for Wikipedia Talk page discussions according to Ferschke (2014).

Segmentation, Classification, and Behavior Vector Construction

On the surface level, Wikipedia Talk pages resemble threaded discussions in web forums. However, they lack the rigid thread structure that is usually enforced by the forum software. Talk pages use the MediaWiki markup for segmenting discussions into individual topics. Furthermore, users are requested to mimic a thread structure by indenting their posts and signing them with their usernames. Since this practice is not enforced by the system, reliably segmenting the unstructured discourse can be a challenging task. We use a variation of the Talk page segmentation algorithm described by Ferschke (2014), which identifies turn boundaries with the help of the Talk page revision history. Furthermore, the approach uses the revision meta data to identify the authors of each turn even if they did not sign their post. In particular, we employ the forward checking approach described by Ferschke and regard each paragraph within a discussion as a separate turn. Users who posted to a discussion while not being logged in are identified by their IP. Even though the same person might theoretically appear with different IPs, the risk of misidentifying individuals is small within the scope of a single discussion. We therefore regard IPs as appropriate identifiers for anonymous individuals for our study.

From the 6,000 selected Talk pages, 2,052 pages had at least one discussion in the observed time period with more than three different participants. Overall, we extracted 5,185 discussions with 72,031 turns posted by 8,042 users. Given the constraint that the articles were observed at an early stage of their development, it is not surprising that only 55 articles in our sample were labeled with the highest quality levels (Good article, A-class, Featured article), while the remaining articles fell in any of the lower quality categories (Stub, Start, B-Class).

We construct the behavior vectors for each user by counting the frequency of each label in all of their posts within a single discussion. We furthermore include the absolute number of posts by that user in this discussion. This results in a 13-dimensional vector for each discussant. For every discussion, the behavior vectors of all participating users are then combined and assigned a quality score based on the article quality rating at the end of the observed time window.

Experiment

In order to increase the likelihood of the role assignments to converge, we initialize our algorithm with a heuristically determined role assignment rather than assigning roles randomly. In our initial experiment, we seek to identify four roles, and assign one individual to each of these roles in each discussion. Based a manual review of Wikipedia Talk pages, we determined which of the dialog act labels frequently co-occur in the posts of individual users indicating that they tend to perform a restricted set of certain acts thus representing a particular role in the discussion. Based on this first intuition, we formed four sets of dialog act labels[2]. We rank the users according to the label frequencies in each category. We then initially assign the user with the highest rank in a category to the respective role.

For each role k, we compute the associated vector representation R_k by averaging all behavior vectors $B_{j,i}$ of participant i in discussion j who are assigned to that role in the final iteration. For each role k and feature i, the regression model furthermore assigns a weight $W_{k,i}$, which measures the importance of the corresponding feature in the respective role for predicting article quality. We can compute an indicator of how much influence a role has for predicting article quality. This indicator is computed by summing over features the product of the average value of the feature in the role and the weight for that feature. The sum over all of these role indicators is the total influence of the model, and the role influence of each role is its indicator value divided by the role weight. We can then interpret each role in terms of its role influence and the set of features with high average value and weight.

Using this approach, we interpreted the model output. The first role learned by the model is the most influential one, accounting for 65% of the total influence. It accounts for more than twice as much influence as the second most important role identified. The average values for almost all features are substantially higher than the corresponding values in all the other roles, indicating that participants assigned this role are highly active in general and contribute a wide variety of types of contributions. However, the ones that account most for its influence are ACTF, CRITCOMPL, CRITSTRUCT, CRITSUIT, and Number of posts. This is a combination of actions that makes sense when one is planning the early stages of a page, wheres accuracy, language and style, and authoritativeness are smaller issues that would apply later. We have termed this role The Doer.

The next most influential roles is one where the influence comes from ATTNEG and CRITAUTH and number of posts. It accounts for 27% of the total influence of the model. We have termed this role The Critiquer. The other behavior pro-

[2]We define the following four sets of labels for the initial role assignments: (CRITCOMPL, CRITACC, CRITAUTH), (CRITLANG, CRITSTRUCT), (ACTF, ACTP), (REQEDIT, REQMAINT)

files also include critiquing behaviors, however this one emphasizes critiquing over other things in terms of doing it frequently and attributing high weight to these acts.

The next role accounts for only 7% of the total influence of the model. That role's influence comes from ATTPOS and number of posts. We have termed this role The Encourager.

The final role, which accounts for the remaining 1% of influence is determined by ACTP, CRITAUTH, CRITSUIT and REQEDIT. The pairing between ACTP and REQEDIT is interesting. Both have relatively high frequency for this role, which suggests that this person may be conceived as The Manager.

Discussion

The first experiment yielded four distinct roles with differential contribution towards the prediction of article quality, namely The Doer, The Critiquer, The Encourager, and The Manager. It is interesting that the behavior profile that is closest to what would intuitively be thought of as a manager role turns out to be so insignificant in the prediction of article quality. On the other hand, the Critiquer could also be considered a sort of manager, who does not explicitly assign work to individuals, but does point out important work that needs to be done. It is interesting to note that while one might expect leadership to play the most prominent role in prediction of article quality, what we see is that it is most essential to have enough workers who are contributing actively, making commitments, and presumably doing a lot of editing work as well.

Conclusion and Current Work

In this paper we have presented a first application of a lightly supervised Role Identification Model (RIM) to data extracted from Wikipedia Talk pages. The objective was to identify roles that are predictive of article quality, and thus might suggest roles that should be represented on a page in order for effective coordination to take place.

In this first experiment, we began by defining role-based behavior profiles in terms of conversation acts defined in earlier work on analysis of Wikipedia Talk pages (Ferschke, Gurevych, and Chebotar 2012). However, though these contribution types are reminiscent of qualitative analyses of role-based behavior in talk pages from earlier work, there are still dimensions of discussion behavior that could be explored for their contribution both positively and negatively to article quality. For example, some earlier work has aimed to characterize the discourse segment purposes or intentions associated with contributions to the Wikipedia Talk pages (Bracewell, Tomlinson, and Wang 2012; Bracewell et al. 2012). And other work characterizing task related contributions in terms of style (Swayamdipta and Rambow 2012). Still other work focuses on strategies like power plays (Bender et al. 2011; Marin, Zhang, and Ostendorf 2011).

In this first experiment, we have identified a set of four roles. However, what we have not explored is the number of each role type that should be present per page. Also, we have not explored the interplay between roles. Both of these are important questions left to answer in our work going forward.

In the future, we will also use the identified role-based behavior profiles to take inventory of which pages are missing key editor roles. These profiles can also be used to identify participants on related pages that might be able to fill those roles where they are missing. Thus, this work might form the foundation for a social recommendation approach to channel coordination effort in Wikipedia where it might have positive impact on quality.

Acknowledgments

The authors gratefully acknowledge Niki Kittur and Bob Kraut for sharing the data used in their earlier work. This work was funded in part by a seedling grant from the Naval Research Lab and funding from Google.

References

Bamman, D.; O'Connor, B.; and Smith, N. A. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 352–361.

Bamman, D.; Underwood, T.; and Smith, N. A. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, 370–379.

Bender, E. M.; Morgan, J. T.; Oxley, M.; Zachry, M.; Hutchinson, B.; Marin, A.; Zhang, B.; and Ostendorf, M. 2011. Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on Languages in Social Media*, 48–57.

Bracewell, D.; Tomlinson, M.; Brunson, M.; Plymale, J.; Bracewell, J.; and Boerger, D. 2012. Annotation of adversarial and collegial social actions in discourse. In *Proceedings of the Sixth Linguistic Annotation Workshop*, 184–192.

Bracewell, D. B.; Tomlinson, M. T.; and Wang, H. 2012. A motif approach for identifying pursuits of power in social discourse. In *Sixth IEEE International Conference on Semantic Computing, ICSC 2012, Palermo, Italy, September 19-21, 2012*, 1–8.

Ferschke, O.; Gurevych, I.; and Chebotar, Y. 2012. Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 777–786.

Ferschke, O.; Zesch, T.; and Gurevych, I. 2011. Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia's Edit History. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations*, 97–102.

Ferschke, O. 2014. *The Quality of Content in Open Online Collaboration Platforms: Approaches to NLP-supported Information Quality Management in Wikipedia*. Ph.D. Dissertation, Technische Universität Darmstadt, Darmstadt.

Forestier, M.; Stavrianou, A.; Velcin, J.; and Zighed, D. A. 2012. Roles in social networks: Methodologies and research issues. *Web Intelli. and Agent Sys.* 10(1):117–133.

Hu, X., and Liu, H. 2012. Social status and role analysis of palin's email network. In *Proceedings of the 21st International Conference on World Wide Web*, 531–532.

Hu, M.; Lim, E.-P.; and Krishnan, R. 2009. Predicting outcome for collaborative featured article nomination in wikipedia. In *Third International AAAI Conference on Weblogs and Social Media*.

Kittur, A., and Kraut, R. E. 2008. Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, 37–46.

Marin, A.; Zhang, B.; and Ostendorf, M. 2011. Detecting forum authority claims in online discussions. In *Proceedings of the Workshop on Languages in Social Media*, 39–47.

Ravindra, K. A.; Magnanti, T. L.; and Orlin, J. B. 1993. *Network flows: theory, algorithms, and applications*. Prentice Hall Englewood Cliffs.

Swayamdipta, S., and Rambow, O. 2012. The pursuit of power and its manifestation in written dialog. In *Sixth IEEE International Conference on Semantic Computing, ICSC 2012, Palermo, Italy, September 19-21, 2012*, 22–29.

Welser, H. T.; Cosley, D.; Kossinets, G.; Lin, A.; Dokshin, F.; Gay, G.; and Smith, M. 2011. Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, 122–129.

Yang, D.; Wen, M.; and Rosé, C. P. (under review). Weakly supervised role identification in teamwork interactions,. In *Proceedings of the 53rd Annual Meeting of the Association for Computional Linguistics*.

Zhao, Y.; Wang, G.; Yu, P. S.; Liu, S.; and Zhang, S. 2013. Inferring social roles and statuses in social networks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 695–703.