Predicting Outcome for Collaborative Featured Article Nomination in Wikipedia

Meiqun Hu and **Ee-Peng Lim**
School of Information Systems
Singapore Management University
80 Stamford Road, Singapore 178902
meiqun.hu@acm.org eplim@smu.edu.sg

Ramayya Krishnan
Heinz College
Carnegie-Mellon University
Pittsburgh, PA 15213 USA
rk2x@cmu.edu

Abstract

In Wikipedia, good articles are wanted. While Wikipedia relies on collaborative effort from online volunteers for quality checking, the process of selecting top quality articles is time consuming. At present, the duty of decision making is shouldered by only a couple of administrators. Aiming to assist in the quality checking cycles so as to cope with the exponential growth of online contributions to Wikipedia, this work studies the task of predicting the outcome of featured article (FA) nominations. We analyze FA candidate (FAC) sessions collected over a period of 3.5 years, and examine the extent to which consensus has been practised in this process. We explore the use of interaction features between FAC reviewers to learn SVM classifiers to predict the nomination outcome. We find that, calibrating the individual user's polarity of opinions as features improves the prediction accuracy significantly.

Introduction

Motivation

Wikipedia is the result of large number of users collaboratively editing articles on a wide range of topics. It is also the most read online encyclopedia today, and has been frequently referenced by Internet users, despite resentment in some academic institutions. The opponents of Wikipedia often cite uneven content quality as the main reason of not approving its use.

In Wikipedia, high quality articles are hence wanted. In addition to collaborative authorship, Wikipedia has designated *featured article* (FA) label for articles representing the best work in Wikipedia. For an article to become *featured*, it has to meet the quality criteria outlined in (Wikipedia 2008b). These criteria cover both content and presentation aspects. Wikipedia users rely on these criteria to judge the quality of articles, and to determine whether to award FA label.

In previous research (Lih 2004; Hu et al. 2007; Stvilia et al. 2008; Druck, Miklau, and McCallum 2008), several models for determining the quality of Wikipedia articles have been proposed and evaluated. While these models seek to use different measures and features to calibrate article quality, they are completely oblivious of the existing workflow that selects featured articles in Wikipedia. In particular, high quality articles do not automatically acquire FA labels. Only articles nominated as *featured article candidate* (FAC) will undergo review by Wikipedia users, who jointly determine if FA label should be awarded. A detailed description of this review process is given in the next section.

In this paper, we analyze FAC nominations generated over a period of more than 3.5 years. We first study the extent to which consensus applies in FAC discussions, and the level of user activities and collaboration. We later address the task of predicting the outcome of featured article nomination. Instead of replacing the existing featured article nomination and review workflow, we seek to understand the article review process so as to supplement it with prediction model. The prediction model will help FAC director to decide whether an article has gone through sufficient deliberation before being awarded FA label. Here, we assume that the articles in nomination are likely to meet some basic quality criteria of FA. Such quality checking can be performed either manually by human nominators, or by heuristic quality models such as those in (Lih 2004; Hu et al. 2007; Stvilia et al. 2008; Druck, Miklau, and McCallum 2008).

As Wikipedia continues to grow, there is an increasing need to have software that automates the two-step process of acquiring FA label. The first step is to select high quality articles and nominate them for FAC review. The second step is to help the FAC director and his delegate deciding whether to award FA label to a nominated article. Quality measurement models that automatically assess the quality of articles in Wikipedia are designed to address the first step in this process. The prediction on the nomination outcome, on the other hand, focuses on the second step.

Objectives and Contributions

Prediction about the outcome of FAC nomination is a new problem that comes with several challenges. Firstly, one observes a multitude of user interactions in the FAC review process, which includes users' commenting, editing and voting activities. It is unclear what features can be derived from the interaction data for learning prediction models. While natural language understanding (NLU) techniques can be used to determine the intent behind the comment text, the accuracy of such techniques is often not very high. In this work, we therefore avoid using NLU techniques. Secondly,

each FAC nomination involves different groups of users who may act differently from users involved in other FAC nominations. The user composition may affect the nomination outcome, but such a hypothesis needs to be carefully verified (Viégas, Wattenberg, and Mckeon 2007).

In this paper, we therefore set off with research objectives as follows, and make the corresponding contributions:

- *To study the interaction data by users generated during FAC nomination and review periods*: We collect an FAC dataset consisting of all featured article nominations (3,196 in total) from January 2004 onwards. The review discussion content of these nominations are also acquired from Wikipedia to provide a rich set of data for the prediction task. The properties of this FAC dataset are analyzed.
- *To predict on the outcome of FAC nominations based on features derived from the review data*: We derive various sets of features from the review sessions of nominated articles, and adopt SVM classifier to predict the outcome of nominations using these sets of features.
- *To evaluate and compare the prediction methods (cum feature sets)*: We evaluate our proposed prediction methods using *area under the curve* (AUC) metric on *precision recall* (PR) curve. Results show that: (i) features that exploit the aggregated voting statistics are most accurate in predicting the outcome if we only predict for closed FAC discussions; (ii) classifiers using active users and discussion features predict more accurately than that using discussion features only; and (iii) classifiers using active users, discussion and collaborator features yield prediction performance comparable to that using active users and discussion features.

Featured Article in Wikipedia

The initiative of identifying high quality articles in Wikipedia has started as early as June 2003. Since then, the label *featured article* (*FA* for short) has been used to refer to these articles, and a small bronze star is used to display the FA status at the top right corner of the page. It was not until early 2004 that the selection process and criteria were formalized. More recently, other forms of *featured content* have been introduced[1], including *featured pictures*, *featured lists*, *featured portals*, *featured topics* and *featured sounds*. In this work, we however focus on featured articles only.

Featured Article Candidate

To acquire FA label, an article must first be nominated as *featured article candidate* (*FAC*). FAC nomination is often followed by a period of discussion by a group of reviewers. During this period, various aspects in the quality of the article are examined, critical improvements are suggested, and more importantly, opinions on whether to promote the article to FA are exchanged. We name an FAC nomination and the discussion that follows collectively as an *FAC session*. Figure 1 depicts the key steps in the process of acquiring FA label.

[1] `http://en.wikipedia.org/wiki/Portal:Featured content`

An FAC session starts when a nomination is raised. The *nominator* gives his or her reason for nominating the article and awaits *comment*s from peer *reviewer*s. Each comment has its commenter's username and a timestamp. A comment may be nested under another comment, indicating that the former responds to the latter. A comment may contain voting phrase(s) that express the reviewer's approval or disapproval of the FA promotion.

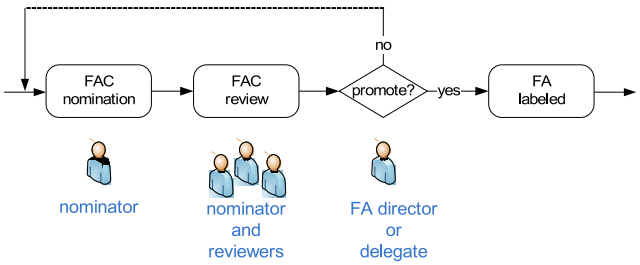


Figure 1: Key steps in acquiring FA label

An FAC session ends when the discussion is closed and a decision can be made on whether an FA label should be awarded. The *FA director* or his/her *delegate* (Wikipedia 2008a) makes the final decisions on when to end the discussion and the outcome of the nomination. FAC sessions usually last for one to two weeks (Wikipedia 2008a), although some articles may require more time to resolve actionable objections.

Each FAC session is archived in Wikipedia. An article may have more than one FAC sessions, if it has been nominated for FA multiple times.

Overview of FAC Dataset

To study the award of FA labels to nominated articles, we crawled all FAC sessions from January 2004 to August 2008. The crawl was done in three key steps. First, we collected the list of articles nominated from month to month. This was done by crawling the page `Wikipedia: Featured article candidates`[2] at the end of each month. Secondly, we located the archived discussion content of each FAC nomination at two sources: (i) URL(s) listed on the page `Wikipedia:Featured article candidates/Featured log`[3] for each month; and (ii) URL(s) shown in the `Article Milestone` section on the `Talk:` page[4] of the article. Lastly, we crawled the archived discussion content of each FAC session, and extracted the outcome of each session from the `Talk:` page of the corresponding article. This gives us 3,196 FAC sessions[5] involving 2,619 articles. Table 1 shows some statis-

[2] This page shows active nominations of the current month.

[3] This page lists all its subpages by month, where each subpage shows the nomination and discussion content of the respective month.

[4] These pages complement the corresponding article pages, where communication among co-authors is carried out.

[5] The content of these FAC sessions will be available upon request from the first author.

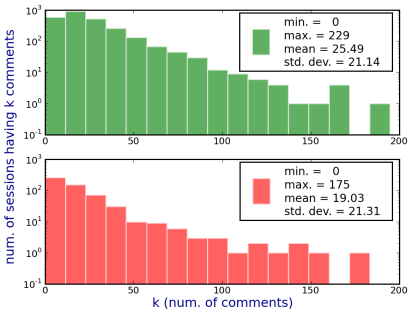


Figure 2: Comments per FAC session

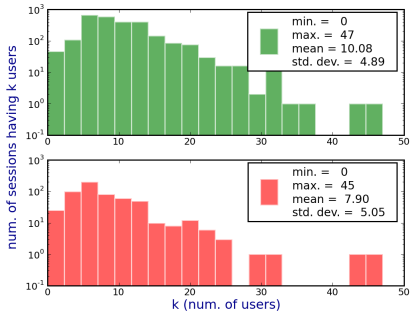


Figure 3: Users per FAC session

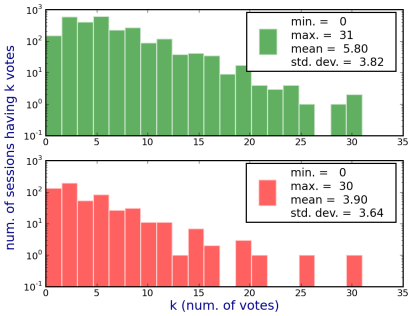


Figure 4: Votes per FAC session

tics about the resulting FAC dataset.

Table 1: Summary statistics of the FAC dataset

num. of articles	2,619	
num. of sessions	3,196	
num. of *passed* sessions	2,633	(82.4%)
num. of *failed* sessions	563	(17.6%)
num. of comments[6]	77,821	
num. of users	4,940	

We observe that there are more *passed* sessions (i.e., sessions that promote the articles) than *failed* sessions (i.e., sessions that fail to promote the articles). This observation also holds for sessions grouped by month. This may due to the requirement that an article should be of high quality before being nominated for FAC (Wikipedia 2008a).

There is an increasing trend in the number of FAC sessions from early months to more recent months. The increasing trend may partly due to broader awareness of FA protocol and the overall increase in Wikipedia's user population and Web traffic. However, the number of FAC sessions grows at a rate that is much lower than the exponential growth of articles in Wikipedia.

The review duration of FAC sessions ranges from 1 to 555 days[7] in our dataset. The average duration is about 11 days.

As shown in Figure 2, the number of comments per session varies from 0 to 229. We observe that most sessions have fewer than 100 comments. On average, the passed sessions have slightly more comments than the failed ones, i.e., 25.49 compared to 19.03.

Figure 3 shows the distribution of the number of distinct users participated in an FAC session. Most sessions have less than 30 users. On average, the passed sessions have slightly more users than the failed ones. This observation is consistent with that for comments and votes in Figure 2 and Figure 4 respectively. Intuitively, the more users participate in the session, the more comments and votes there would be.

[6]This excludes comments for which the commenter cannot be identified.

[7]The FAC nomination for article `Speech synthesis` initiated in May 2004 has lasted for 555 days. However, all comments but the last one were given during May 2004. The actual ending date is subject to verification.

Dissecting the Dataset

In this section, we take a deeper look into three aspects of the FAC dataset: (i) consensus in FAC sessions; (ii) user activeness; and (iii) users' collaborative relationship.

Consensus in FAC Discussion

Consensus implies majority agreement. Wikipedia advocates the use of consensus to determine the outcome of a nomination. As part of comment writing, an FAC reviewer may cast his/her *vote*, in the form of `support` or `objection` (equivalently `oppose`), to express individual's opinion on whether to award FA label to the article. Reaching consensus, while an ideal principle, is by no means easy in practice. We therefore study the extent to which the consensus principle has been adopted by examining the FAC dataset.

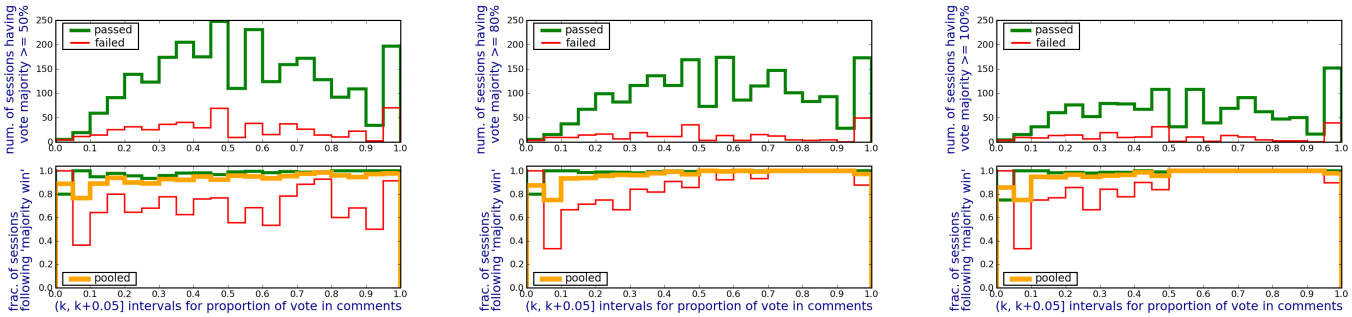
We first examine the number of votes in FAC sessions, shown in Figure 4. It shows that most sessions have fewer than 30 votes. A passed session has on average 7.42 votes while a failed session has on average 5.11 votes. Compared to the number of comments per session (Figure 2), votes are much fewer, suggesting that not many users perform voting. This is possibly due to: (i) many comments are written responding to other comments, rather than directly to respond to the nomination; (ii) many comments that do respond to the nomination do not express approval or disapproval.

The principle of *consensus* requires a session to have high proportion of vote and the *majority* should win. We define the *proportion of vote* in a session by

$$\frac{\text{number of comments that contain voting phrase(s)}}{\text{number of comments that respond to the nomination}}$$

Clearly, this proportion falls in the range $[0, 1]$. We divide the spectrum of proportion of vote into 20 intervals, each with a width of 0.05. We examine the extent to which the principle of *majority win* is followed by sessions in each interval of proportion of vote.

Given that an FAC session consists of votes from multiple users, we call a session following the principle of majority win if its final outcome is consistent with the majority reviewers who has voted. Intuitively, it is expected that the principle should be followed by most of the sessions. However, different thresholds can be adopted in determining *majority*. A majority of $t\%$ means that more than $t\%$ of all voters in the session hold to one opinion (either approval



(a) Majority $\geq 50\%$ (b) Majority $\geq 80\%$ (c) Majority $\geq 100\%$

Figure 5: (Best viewed in color.) Consensus in FAC discussions

or disapproval) and the remaining $(100 - t)\%$ voters vote otherwise. In Figure 5, we vary this threshold from 50% to 100%, and plot the number of sessions satisfying the threshold (upper subplot) and the fraction of sessions, among those that pass the threshold, also following the principle of majority win (lower subplot). The additional bold line in the lower subplot shows the pooled fraction from both passed and failed sessions.

It is interesting to note that few sessions have very large proportion of vote. The density distributions (visually) follow normal distribution. We find the modes mostly lie in the range from 0.45 to 0.60. Moreover, for the interval with the highest proportion of vote, i.e., $(0.95, 1.0]$, the fraction of majority for these sessions remains high regardless of the threshold.

As shown in Figure 5(a), when using 50% as the threshold for majority, the principle of majority win works very well for the passed sessions but less so for the failed sessions. When we increase the threshold for majority to 80%, failed sessions are found to follow the principle, as shown in Figure 5(b). Only the failed sessions with small proportion of vote, i.e., intervals $(0.00, 0.30]$, are found to defy this principle. This is reasonable considering that FA director may decide not to follow the votes if there are too few votes or voters.

It appears that consensus is a good principle for deciding the nomination outcome. However, in view that votes are not always made in an FAC session, and votes as direct features themselves are not always available from all reviewers, we have considered the use of other non-voting features to predict the outcome. We elaborate these features in the next section.

To summarize what we observe so far:

- Reviewers do not always vote in an FAC session. There are FAC sessions with very low proportion of vote.
- Not all FAC sessions achieve consensus, i.e., high proportion of vote and majority win, on the nomination outcome. When consensus is reached, it is most likely that the final nomination outcome is consistent with the opinion of the majority voters. This observation is more prevalent among the passed sessions than the failed sessions. The threshold for majority at 80% appears to be reasonable.

User Activeness

Users of Wikipedia demonstrate different levels of activeness when they participate in FAC discussion. A user is said to be highly *active* if he/she participates in a large number of FAC-related activity. In this section, we examine users' *activeness* using metrics such as the number of nominations, the number of sessions and the number of comments.

Figure 6 summarizes users using these metrics. It shows that most users do not nominate many FA candidates, neither do they participate in many FAC sessions or give many comments. For each respective metric, there are only a handful of users who are highly active in our dataset. Among all FAC sessions, there are only $1,272$ distinct nominators and $4,849$ distinct reviewers who comment on nominations.

Next, we examine how correlated are the three metrics of user activeness. Table 2 below shows that, there is a strong positive correlation between the number of comments and the number of sessions among the users. These two metrics, however, are weakly correlated with the number of FAC nominations.

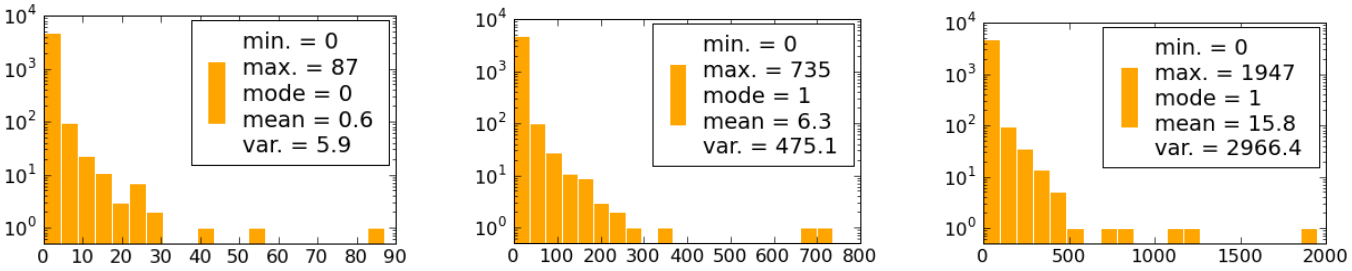
Table 2: Pearson correlation coefficients between three metrics on users' activeness

	# Nom	# Ses	# Com
# Nomination	1	0.546	0.469
# Sessions	-	1	0.921
# Comments	-	-	1

Collaborative Relationship between Users

We also consider the relationship between pairs of users participating in FAC sessions. Since there are no explicit inter-user relationships provided in Wikipedia, we examine the relationships formed through co-reviewing an FAC session by each pair of users.

Out of the $4,849$ distinct users, $101,845$ pairs co-review common sessions. Only very few pairs co-review (relatively) large number of common sessions, i.e., 438 pairs co-reviewed 15 sessions or more, and 91 pairs co-reviewed 30 sessions or more. These strong *ties* may appear less accidental than others. They represent strong collaborative relationships among users as they work together in criticizing



(a) Nomination (b) Session (c) Comment

Figure 6: Distribution of users' activeness in three metrics

and improving articles. We later exploit such pair features in predicting the outcome of FAC sessions.

Predicting FAC Outcome

We cast the FAC outcome prediction task into a binary classification problem, where each session instance is represented using a set of features and the likelihood of each instance being positive (*pass*) or negative (*fail*) is to be predicted by the classifier.

Feature Engineering

We first identify features that can be relevant to predicting the nomination outcome. We divide them into three categories, namely **discussion features**, **user features**, and **collaborator features**.

Discussion Features are extracted from only the text content of each FAC session. As noted, passed and failed sessions show differences in distributions of session duration, the number of comments and the number of distinct users. Hence, we consider the following *general discussion features*: (1) *duration (in days) of the session*, (2) *total number of comments* (excluding the nomination comment), (3) *total number of distinct users*, and (4) *average number of comments per user*.

We also derive *comment specific discussion features*: (5-6) *maximum* and *average length of comments*, (7-8) *maximum* and *average depth of comments*. The *depth of comment* refers to the level at which the comment is nested under other comment(s). For example, a comment that directly responds to the nomination is at depth 1; if a comment responds to another comment at depth 1, the former is at depth 2; and so forth. We expect these comment specific discussion features to reflect the deliberation structure among reviewers.

To consider the participation of the nominator, as well as FA director and delegate in an FAC session, we also include three binary features: (9) *self nomination* (i.e., the nomination is raised by a user who also contributes to the article); (10) *director commentation* (i.e., the FA director `User:Raul654` participates in the session; (11) *director's delegate commentation* (i.e., the director's delegate `User:SandyGeorgia` participates in the session).

None of the above discussion features is related to voting thus far. The *voting specific discussion features* are those that are derived from users' opinion on approving or disapproving the nomination: (12) *number of comments at depth 1* (i.e., these are comments that directly respond to the nomination); (13) *number of voting comments* (i.e., comments at depth 1 that also contain voting phrase(s)); (14) *fraction of comments that vote for support*; and (15) *fraction of comments that vote for objection*.

User Features are the set of features that are defined on the user dimensions. As shown in Figure 6, users exhibit different levels of activeness in the FAC sessions. User features allow us to examine the hypothesis that "active users are more influential than the less active ones for predicting FAC nomination". If the hypothesis holds, user features may help to improve prediction accuracy. Nevertheless, to define user features, two questions need to be answered: (i) how to select the active users? and (ii) what user features can be defined for each FAC session?

To identify active users, we select top 50 users (slightly over 1% of distinct users found in our FAC dataset) ranked by: (i) the number of nominations he/she raises (N); (ii) the number of FAC sessions he/she participates (S); (iii) the number of comments he/she contributes in all FAC sessions (C); and (iv) the number of distinct co-reviewer links he/she forms (L).

Table 3 shows the Jaccard coefficient between choices of top 50 users selected by the different metrics of activeness. Jaccard coefficient of two sets of active users, U_1 and U_2, is defined by

$$J(U_1, U_2) = \frac{|U_1 \cap U_2|}{|U_1 \cup U_2|}$$

in which values fall between 0 (completely non-overlapping) and 1 (identical). As shown in Table 3, the top 50 users by the number of co-reviewer links are the most similar to those by the number of sessions, i.e., $J(U_L, U_S) = 0.695$.

Based on the selected active users, we define user features as follows:

- User existence ($\mathsf{e_u}$): The feature value is 1 when the corresponding active user participates in the session, and 0 otherwise.
- User comment count ($\mathsf{c_u}$): The number of comments given in the session by the corresponding active user.
- User vote polarity ($\mathsf{p_u}$): The feature value is 1 if the corresponding active user is present in the session and votes

Table 3: Jaccard coefficient between top 50 users selected by four metrics of activeness

	Sessions covered	Jaccard coefficient		
		N	S	C
Nomination	2,589	-		
Session	2,882	0.266	-	
Comment	2,794	0.351	0.538	-
Co-session link	2,864	0.250	0.695	0.429

for *support*, and it is -1 if the user is present but votes for *objection*, and 0 otherwise (i.e., either the user does not participate in the session or does not vote). This vote polarity feature is different from the consensus measure or voting specific features discussed earlier. It is derived for the individual user and does not involve aggregated voting statistics.

- Signed comment count ($\mathsf{s_u}$): The product of feature values from $\mathsf{p_u}$ and $\mathsf{c_u}$.

Collaborator Features are defined on the dimension of pairs of users. Again, we are interested in top pairs based on user collaboration. The metrics for selecting the most collaborative pairs are: (i) the number of FAC sessions the two users co-reviewed (Co); (ii) the degree to which the two users agree in their co-reviewed sessions (Ag); (iii) the degree to which the two users disagree (Dg). We define the *degree of agreement* (and similarly for disagreement) between a pair of users by

$$\frac{\text{number of sessions the pair agree in their votes}}{\text{number of sessions the pair both voted}}$$

Co is an obvious choice when considering the collaboration between pairs of users, while the other two metrics account for two extreme scenarios. Ag helps us to identify pairs of co-reviewers that mostly supporting each other. On the other hand, Dg finds the pairs that most often hold opposite opinions. The choice of latter is intended for exposing sessions that mostly contain words from both sides. We compare the choice of these three metrics in experiments.

Table 4: Jaccard coefficient between top 100 pairs of users by metrics of collaboration

	Sessions covered	Jaccard coefficient	
		Co	Ag
Co-session	1,501	-	
Agreement	1,184	0.136	-
Disagreement	834	0.000	0.000

Unlike top active users, the top pairs determined by different metrics of collaboration are present in less than 50% of all FAC sessions. These pairs are quite distinct, since the Jaccard coefficients between sets of top pairs are small, as shown in Table 4.

Based on the top pairs of users, we define feature values on the collaborator dimensions as follows:

- User pair existence ($\mathsf{e_p}$): The feature value is 1 if both users participate in the session, and 0 otherwise.
- User pair comments ($\mathsf{c_p}$): Sum of the number of comments contributed to the session by the two users.
- User pair polarity - option 1 ($\mathsf{p1_p}$): Sum of the polarity of the two users. The polarity value of an individual user is defined the same as in $\mathsf{p_u}$. The possible values for sum of polarities from two users are {-2,-1,0,1,2}. Clearly, this assignment cannot distinguish the cases of $\langle +1, -1\rangle$ from those of $\langle 0, 0\rangle$. To address this, we have $\mathsf{p2_p}$.
- User pair polarity - option 2 ($\mathsf{p2_p}$): Sum of the polarity of the two users, except that -0.5 is assigned when one of the pair votes for objection and the other votes for support. It is intended to distinguish the case where objection is perceived to be more severe than support. Hence, the possible feature values are {-2,-1,-0.5,0,1,2}.

Prediction Methods

We encode each feature setting by a triple

$$\langle \mathsf{D_x}, \mathsf{U_y}(\mathsf{F_u}), \mathsf{P_z}(\mathsf{F_p})\rangle$$

where,

- $\mathsf{D_x}$ denotes the set of discussion features, where $\mathsf{x} \subseteq \{\mathsf{g}+\mathsf{c}, \mathsf{v}\}$ with $\mathsf{g}+\mathsf{c}$ refers to general and comment specific discussion features, and v refers to voting specific discussion features.
- $\mathsf{U_y}(\mathsf{F_u})$ denotes the set of user features defined based on top active users. Here, $\mathsf{y} \in \{\mathsf{N}, \mathsf{S}, \mathsf{C}, \mathsf{L}\}$ refers to the set of top active users selected by the corresponding metric of activeness, $\mathsf{F_u} \in \{\mathsf{e_u}, \mathsf{c_u}, \mathsf{p_u}, \mathsf{s_u}\}$ refers to the option in assigning feature values on user dimensions.
- $\mathsf{P_z}(\mathsf{F_p})$ denotes the set of collaborator features defined based on top pairs of collaborating users, in which pairs are selected by one of the collaboration metric, i.e., $\mathsf{z} \in \{\mathsf{Co}, \mathsf{Ag}, \mathsf{Dg}\}$, and feature values on pair dimensions are determined by one of the options $\mathsf{F_p} \in \{\mathsf{e_p}, \mathsf{c_p}, \mathsf{p1_p}, \mathsf{p2_p}\}$.

For each chosen feature setting, we train a SVM classifier and evaluate its classification performance. We use linear kernel for SVM, since linear kernel enables us to find and interpret the separating hyperplane determined by the classifier.

Evaluation and Results

For performance comparison, we adopt *area under the curve* (AUC) metric on *precision-recall* curve (PR curve). PR curve is more suitable than ROC (*receiver operation characteristic*) curve for comparison on imbalanced dataset (Davis and Goadrich 2006). Our FAC dataset is imbalanced, since 82.38% instances are positive and 17.62% are negative. Precision and recall are computed for the negative class, since negative instances are the minority.

We partition our FAC dataset into 10 folds, using stratified sampling based on outcome. We use SVM^{light} (Joachims 1999) for learning and classification. A cost factor of 0.2 is used in learning, which is derived from the ratio between

negative and positive instances, i.e., $\frac{n_-}{n_+} = \frac{563}{2633} \simeq 0.219$. We apply standardization (*Z-normalization*) on feature dimensions that are not binary, since it is noted to achieve faster convergence in SVM^{light}. Finally, we adopt Platt's calibration method (Lin, Lin, and Weng 2007) to calibrate SVM decision values into class posterior probabilities.

Using Discussion Features Table 5 shows the average AUC (over 10 folds) on PR curve given by SVM classifiers using discussion features only.

Table 5: AUC (on PR) using discussion features

$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \emptyset \rangle$	0.402 ($\pm$0.063)
$\langle \mathsf{D}_{\{\mathsf{v}\}}, \emptyset, \emptyset \rangle$	0.816 ($\pm$0.057)
$\langle \mathsf{D}_{\{\mathsf{g+c,v}\}}, \emptyset, \emptyset \rangle$	0.822 ($\pm$0.052)
baseline	0.176

We observe that using voting specific discussion features ($\langle \mathsf{D}_{\{\mathsf{v}\}}, \emptyset, \emptyset \rangle$) outperforms that of using non-voting discussion features ($\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \emptyset \rangle$). The setting $\langle \mathsf{D}_{\{\mathsf{g+c,v}\}}, \emptyset, \emptyset \rangle$, which consists of both voting specific and non-voting discussion features, performs better than the rest. All $\mathsf{D}_\times$ settings outperform the baseline, which is the maximum prior classifier.

On the whole, we could confirm our expectation that, voting specific discussion features are effective in predicting the nomination outcome. However, these features, mostly aggregated voting statistics, are available only when the FAC review is about to end. Given that the time duration of FAC review varies widely and such voting statistics fluctuate throughout the period, it imposes challenge on using these features on a timely basis. In our attempt of using user features ($\mathsf{U}_\mathsf{y}(\mathsf{F}_\mathsf{u})$) and collaborator features ($\mathsf{P}_\mathsf{z}(\mathsf{F}_\mathsf{p})$) in addition to non-voting discussion features ($\mathsf{D}_{\{\mathsf{g+c}\}}$), such time constraint is avoided, yet we could achieve improved prediction accuracy, as shown next.

Using User Features Table 6 shows the average AUC on PR curve when user features are used in addition to non-voting discussion features[8].

On the whole, adding user features is superior over using only non-voting discussion features. This is largely confirmed by 14 out of 16 feature settings on $\mathsf{U}_\mathsf{y}(\mathsf{F}_\mathsf{u})$, as our paired one-tail *t*-tests show significant improvement with significance level of 95%.

We also notice that, p_u always outperforms s_u, and e_u always outperforms c_u, regardless the choice of U_y. Note that c_u (s_u) amplifies e_u (p_u respectively) by a magnitude that is the number of comments given the user in the session. Shown in Table 6, such amplification hurts prediction accuracy. This result suggests, the number of comments is not critical in this prediction task. To understand this result, we find, in most cases users give more than one comment in an FAC session mainly to respond to other reviewer(s). These users may be either the nominator or main contributor(s) of the article. Assuming that users do not switch opinion (approving or disapproving the nomination) during the review period, it makes sense that one comment (mostly responds to the nomination directly but not nested under other comment(s)) is sufficient. The additional comments mainly serve to support their opinions stated earlier. Therefore, it is not a surprise to see p_u outperforms s_u.

Table 6: AUC (on PR) using user features

$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{N}(\mathsf{e}_\mathsf{u}), \emptyset \rangle$	0.438* ($\pm$0.060)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{N}(\mathsf{c}_\mathsf{u}), \emptyset \rangle$	0.432* ($\pm$0.071)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{N}(\mathsf{p}_\mathsf{u}), \emptyset \rangle$	0.511* ($\pm$0.068)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{N}(\mathsf{s}_\mathsf{u}), \emptyset \rangle$	0.468* ($\pm$0.067)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{S}(\mathsf{e}_\mathsf{u}), \emptyset \rangle$	0.439* ($\pm$0.064)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{S}(\mathsf{c}_\mathsf{u}), \emptyset \rangle$	0.413 ($\pm$0.057)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{S}(\mathsf{p}_\mathsf{u}), \emptyset \rangle$	0.590* ($\pm$0.052)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{S}(\mathsf{s}_\mathsf{u}), \emptyset \rangle$	0.470* ($\pm$0.062)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{C}(\mathsf{e}_\mathsf{u}), \emptyset \rangle$	0.446* ($\pm$0.051)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{C}(\mathsf{c}_\mathsf{u}), \emptyset \rangle$	0.429* ($\pm$0.055)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{C}(\mathsf{p}_\mathsf{u}), \emptyset \rangle$	0.558* ($\pm$0.050)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{C}(\mathsf{s}_\mathsf{u}), \emptyset \rangle$	0.460* ($\pm$0.070)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{L}(\mathsf{e}_\mathsf{u}), \emptyset \rangle$	0.440* ($\pm$0.063)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{L}(\mathsf{c}_\mathsf{u}), \emptyset \rangle$	0.406 ($\pm$0.056)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{L}(\mathsf{p}_\mathsf{u}), \emptyset \rangle$	0.586* ($\pm$0.055)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \mathsf{U}_\mathsf{L}(\mathsf{s}_\mathsf{u}), \emptyset \rangle$	0.469* ($\pm$0.062)

Using Collaborator Features The average AUC on PR curve using collaborator features is shown in Table 7.

Table 7: AUC (on PR) using collaborator features

$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \mathsf{P}_{\mathsf{Co}}(\mathsf{e}_\mathsf{p}) \rangle$	0.383 ($\pm$0.058)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \mathsf{P}_{\mathsf{Co}}(\mathsf{c}_\mathsf{p}) \rangle$	0.369 ($\pm$0.054)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \mathsf{P}_{\mathsf{Co}}(\mathsf{p1}_\mathsf{p}) \rangle$	0.556* ($\pm$0.037)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \mathsf{P}_{\mathsf{Co}}(\mathsf{p2}_\mathsf{p}) \rangle$	0.552* ($\pm$0.032)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \mathsf{P}_{\mathsf{Ag}}(\mathsf{e}_\mathsf{p}) \rangle$	0.397 ($\pm$0.043)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \mathsf{P}_{\mathsf{Ag}}(\mathsf{c}_\mathsf{p}) \rangle$	0.388 ($\pm$0.061)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \mathsf{P}_{\mathsf{Ag}}(\mathsf{p1}_\mathsf{p}) \rangle$	0.571* ($\pm$0.067)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \mathsf{P}_{\mathsf{Ag}}(\mathsf{p2}_\mathsf{p}) \rangle$	0.572* ($\pm$0.067)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \mathsf{P}_{\mathsf{Dg}}(\mathsf{e}_\mathsf{p}) \rangle$	0.375 ($\pm$0.053)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \mathsf{P}_{\mathsf{Dg}}(\mathsf{c}_\mathsf{p}) \rangle$	0.377 ($\pm$0.062)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \mathsf{P}_{\mathsf{Dg}}(\mathsf{p1}_\mathsf{p}) \rangle$	0.568* ($\pm$0.075)
$\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \mathsf{P}_{\mathsf{Dg}}(\mathsf{p2}_\mathsf{p}) \rangle$	0.560* ($\pm$0.067)

It is not a surprise to see settings using $\mathsf{p1}_\mathsf{p}$ and $\mathsf{p2}_\mathsf{p}$ improve AUC when collaborator features are used in addition to non-voting discussion features. It suggests that the combined opinion from pairs of users still play a big role in predicting the nomination outcome. This is consistent with Table 6, but only on individual users. Unfortunately, settings using e_p and c_p fail to improve the prediction performance.

Out of our expectation, $\mathsf{p1}_\mathsf{p}$ and $\mathsf{p2}_\mathsf{p}$ do not differ from each other significantly in AUC performance. The difference between options $\mathsf{p1}_\mathsf{p}$ and $\mathsf{p2}_\mathsf{p}$ only happens when both users are present in the session and hold opposite opin-

[8] * denoted settings that perform significantly better than $\langle \mathsf{D}_{\{\mathsf{g+c}\}}, \emptyset, \emptyset \rangle$, based on paired one-tail *t*-test with significance level of 95%.

ions. In those cases, we take objection more severe than support in $p2_p$, whereas we treat opinions from both side equally in $p1_p$. Therefore, we expect the largest difference in AUC between $p1_p$ and $p2_p$ be shown in the settings of P_{Dg}. This is confirmed by the last two rows of Table 7, as compared to other P_z settings. However, letting the opposer have more say is not as good as taking both the supporter and the opposer equally. Nonetheless, $0.568\,(\pm 0.075)$ by $P_{Dg}(p1_p)$ is not significantly better than $0.560\,(\pm 0.067)$ given by $P_{Dg}(p2_p)$.

Using the 'Best of Bests' Settings Lastly, we pick the top two performing $U_y(F_u)$ settings and top two performing $P_z(F_p)$ settings, and merge them to form four new feature settings. For user features, $U_S(p_u)$ and $U_L(p_u)$ give competitively good AUC performance. For collaborator features, we choose $p1_p$ and $p2_p$ each in combination with P_{Ag}. The resulting triples are
$\langle D_{\{g+c\}}, U_S(p_u), P_{Ag}(p1_p)\rangle$, $\langle D_{\{g+c\}}, U_S(p_u), P_{Ag}(p2_p)\rangle$,
$\langle D_{\{g+c\}}, U_L(p_u), P_{Ag}(p1_p)\rangle$, $\langle D_{\{g+c\}}, U_L(p_u), P_{Ag}(p2_p)\rangle$.
We show their AUC performance given by linear SVM classifier in Table 8[9].

Table 8: AUC (on PR) using the 'best of bests' features

$\langle D_{\{g+c\}}, U_S(p_u), P_{Ag}(p1_p)\rangle$	0.593 (± 0.069)
$\langle D_{\{g+c\}}, U_S(p_u), P_{Ag}(p2_p)\rangle$	0.592 (± 0.069)
$\langle D_{\{g+c\}}, U_L(p_u), P_{Ag}(p1_p)\rangle$	0.598 (± 0.069)
$\langle D_{\{g+c\}}, U_L(p_u), P_{Ag}(p2_p)\rangle$	0.598 (± 0.070)
$\langle D_{\{g+c\}}, U_S(p_u), \emptyset \rangle$	0.590 (± 0.052)
$\langle D_{\{g+c\}}, U_L(p_u), \emptyset \rangle$	0.586 (± 0.055)
$\langle D_{\{g+c\}}, \emptyset, P_{Ag}(p1_p)\rangle$	0.571 (± 0.067)
$\langle D_{\{g+c\}}, \emptyset, P_{Ag}(p2_p)\rangle$	0.572 (± 0.067)

It is expected that, for linear SVM, using more features gives better performance than using smaller subset of features. This expectation is largely confirmed by the average AUC shown in Table 8. Moreover, adding user features to $\langle D_{\{g+c\}}, \emptyset, P_{Ag}(F_p)\rangle$ settings always improves AUC performance significantly, as suggested by our paired one-tail *t*-tests with significance level of 95%. On the contrary, when adding collaborator features to $\langle D_{\{g+c\}}, U_S(p_u), \emptyset\rangle$ and $\langle D_{\{g+c\}}, U_L(p_u), \emptyset\rangle$ settings, the increment in AUC is not statistically significant.

Conclusion

Nomination of featured articles in Wikipedia is a process of collaboration. This paper analyzes user collaborations in the nomination of featured articles, by constructing a unique set of featured article candidate (FAC) dataset. We examine users' participation, commenting and voting statistics in the dataset, as well as the adoption of consensus as the decision making principle. We also address the prediction on the nomination outcome as a binary classification task, where features involving discussion content, active users, and collaborative user pairs are identified for each FAC session. Using SVM classifiers, we show that the prediction performance using user features in addition to discussion features is significantly better than using discussion features only. On the other hand, collaborator features do not show significantly in improving the prediction.

Community coordinated decision making has been widely adopted in Wikipedia. One other example is *Requests for Adminship*, in which the community decides which user (upon request) would become administrators. This work represents one of the first kind in predicting outcomes of user collaboration, and there is much room for future research. In particular, we plan to study the connectivity among users through collaboration which may reveal interesting patterns that help determining the nomination outcome even more accurately.

[9]The top two performing $U_y(F_u)$ settings and top two performing $P_z(F_p)$ settings are included at the bottom of the table for easy reference.

Acknowledgement

We acknowledge the partial support for this work from research grant NRF2008IDM-IDM004-036.

References

Davis, J., and Goadrich, M. 2006. The relationship between precision-recall and ROC curves. In *Proceedings of ICML'06*, 233–240. ACM.

Druck, G.; Miklau, G.; and McCallum, A. 2008. Learning to predict the quality of contributions to Wikipedia. In *Proceedings of AAAI'08 Workshop on Wikipedia and Artificial Intelligence*, 983–1001. AAAI Press.

Hu, M.; Lim, E.-P.; Sun, A.; Lauw, H. W.; and Vuong, B.-Q. 2007. Measuring article quality in Wikipedia: models and evaluation. In *Proceedings of CIKM'07*, 243–252. ACM.

Joachims, T. 1999. Making large-scale SVM learning practical. In Schölkopf, B.; Burges, C. J. C.; and Smola, A. J., eds., *Advances in Kernel Methods - Support Vector Learning*. MIT Press. 169–184.

Lih, A. 2004. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In *Proceedings of the Fifth International Symposium on Online Journalism*.

Lin, H.-T.; Lin, C.-J.; and Weng, R. 2007. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning* 68(3):267–276.

Stvilia, B.; Twidale, M. B.; Smith, L. C.; and Gasser, L. 2008. Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology* 59(6):983–1001.

Viégas, F. B.; Wattenberg, M.; and Mckeon, M. 2007. The hidden order of Wikipedia. In *Online Communities and Social Computing*. 445–454.

Wikipedia. 2008a. Featured article candidates. `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_candidates`.

Wikipedia. 2008b. Featured article criteria. `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria`.