Leveraging Editor Collaboration Patterns in Wikipedia

Hoda Sepehri Rad Aibek Makazhanov Davood Rafiei Denilson Barbosa
Department of Computing Science
University of Alberta,Edmonton,Canada
{sepehrir,makazhan,drafiei,denilson}@ualberta.ca

ABSTRACT

Predicting the positive or negative attitude of individuals towards each other in a social environment has long been of interest, with applications in many domains. We investigate this problem in the context of the collaborative editing of articles in Wikipedia, showing that there is enough information in the edit history of the articles that can be utilized for predicting the attitude of co-editors. We train a model using a distant supervision approach, by labeling interactions between editors as positive or negative depending on how these editors vote for each other in Wikipedia admin elections. We use the model to predict the attitude among other editors, who have neither run nor voted in an election. We validate our model by assessing its accuracy in the tasks of predicting the results of the actual elections, and identifying controversial articles. Our analysis reveals that the interactions in co-editing articles can accurately predict votes, although there are differences between positive and negative votes. For instance, the accuracy when predicting negative votes substantially increases by considering longer traces of the edit history. As for predicting controversial articles, we show that exploiting positive and negative interactions during the production of an article provides substantial improvements on previous attempts at detecting controversial articles in Wikipedia.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Wikipedia, Admin Election, Social Interactions, Controversial Articles

1. INTRODUCTION

The recent proliferation of social technologies such as Weblogs, Wikis, social networking sites, etc. has been met with widespread adoption by a large fraction of the general population, who have crossed the line from consumers to producers of content. This phenomenon, which now seems irreversible, has had a tremendous impact on how large segments of society get informed and educated. One particular example of this trend is Wikipedia, which has become one of the 5 most visited websites [4] (up from 500-th in 2004). Anecdotal evidence points to problems and virtues of relying on Wikipedia [9], but the trend seems to be that Wikipedia will indeed become the primary source of reference for most common knowledge in the world.

time	editor	action	Δ	comment
3:55	Infinity0	Rv	—	revert weasel words and pov
4:06	RJII	Rv	412	revert to rjii infinity is misleading the readers to think that tucker opposes employee employer relations...
4:09	Infinity0	Rv	-412	it says that tucker supported private mop please read your version uses many weasel words
4:12	RJII	Del	-131	anarcho capitalism tag
4:15	RJII	Ins	382	noting that tucker supports liberty of people to engage in employee employer relationships don't censor this fact
4:29	Infinity0	Del	-12	anarcho capitalism what's dubious it's a direct quote
5:21	Infinity0	Del	-264	anarcho capitalism
12:03	*other*†	Ins	41	ruined it

† Different user, with id VolatileChemical.

Figure 1: Partial edit history of article on Anarchism.

One of the major strengths of Wikipedia, its openness, is also one of its main sources of criticism. The argument is that virtually anyone can edit any Wikipedia article, regardless of their intentions and/or knowledge about the topic of said article. Even for articles where all editors are well intentioned and knowledgeable, the diversity of opinions and points of view will lead to disagreement during the editing process. Since editing is a continuous process, it is possible that consecutive visits to the same article return drastically different material, possibly reflecting the different biases and opinions of different editors. Quality control in this model is delegated to the crowd—if the topic is important to a large enough group of editors, the collaborative editing process will (eventually) lead to a high-quality article. Also, the entire *edit history* of the article is made available to the reader—who could in principle inspect it before deciding whether or not to trust the content (see, e.g., Figure 1). Finally, articles whose editing process deteriorate into flagrant disputes are explicitly marked as controversial by a group

HT'12, June 25–28, 2012, Milwaukee, Wisconsin, USA.

of *administrators*, most of whom are elected by their peers (other editors), warning the readers about the contentious or disputable nature of the content that they may read.

A question that arises in this context is whether or not one can detect the attitude (positive or negative) of editors towards each other from the recorded history of interactions between them, in the articles they co-edit. This would be useful to automate the identification of controversial articles, for instance. Figure 1 shows a small fragment of the edit history of the article on Anarchism around March of 2006 (the "Δ" column indicates the net change in length of the article, measured in characters between consecutive versions). At the time of writing, the article contained 15,767 edits. We focus on the interactions between two editors: *RJII*, who contributed 1,544 edits, and *Infinity0*, who made 433 edits. The disagreement between these editors is evidenced by their direct mutual accusations and the difference in their use of language: in this article, on average, *RJII* writes longer comments than *Infinity0* (70.6 characters vs 49.3 characters) and also uses more *positive* terms in his comments (423 versus 115). The sequence and timing of the actions is also revealing. The two editors are working concurrently, sometimes *fully* undoing each other's work (called *reverting* versions and indicated as Rv or revert action in the Wikipedia logs) and other times doing so *partially*, by deleting or inserting content to the previous version (indicated as Del and Ins actions, respectively).

While the history snippet above is clear evidence that these two editors did not collaborate, it should be clear that analyzing the edit history of articles in search of disagreements and potential controversies would be virtually impossible for the reader. The sheer volume of data and the frequency with which the edit histories change make such an approach impractical. Moreover, not every editor writes descriptive comments. In fact, there were several examples in further interactions involving *Infinity0* in which he/she would simply revert back to a previous version without any justification. Another issue is that focusing on individual editors is unlikely to lead to good results as well. Again in this example, several editors "teamed up" with *RJII* against *Infinity0*.

1.1 Our Contribution

We propose a method to predict the attitude (positive or negative) between two editors based on the edit history of their interactions. Using this method, we build a signed network [23] of all editors of an article which allows us to infer whether or not the said article contained controversial material.

Our methods are based on machine learning, using classifiers. To obtain labeled training data, we resort to a *distant supervision* approach, using the *records* of Wikipedia administrator elections. For every vote of editor e_1 for editor e_2, we extract all interactions between them and label such interactions according to the vote. Our intuition is that a vote for administrator is an unequivocal declaration of *agreement* or *disagreement* among editors. Indeed, in our motivating example, *RJII* voted against *Infinity0* when he ran for administrator, emphatically voicing his opinion with comments such as "NO WAY! The kid is OUT OF CONTROL... [his] philosophy is to ban an editor whose edits would otherwise prevail... EXTREMELY unethical".

We validate our model in two ways. First we use it to predict actual votes in the administrator elections. Our results show an interesting contrast: overall, one can predict *positive votes* among editors with a markedly higher accuracy. However, the accuracy when predicting negative votes increases more as we increase the number of interactions in the edit history. This suggests that positive interactions and attitudes between editors are the norm, and that negative interactions are not easily forgotten by editors. We also validate our classifier for controversial content on a sample of 480 articles (240 marked as controversial by the administrators). Our results are very encouraging, leading to 84% accuracy overall. We compare our method against other alternatives, including a classifier based on a completely different set of features, and whose accuracy is around 75%. We combine both classifiers as well, increasing the accuracy further to almost 90% in our tests.

2. RELATED WORK

Our work relates to the following areas.

Trust management in Wikipedia. Among the large body of work on Wikipedia, our work mostly relates to trust and reputation management, where a trust score is assigned to an article [6,16,34], to selected parts of an article [2], or even to contributors [3,7,17]. These works often use information from the edit history (or the so-called revision history) of an article, including edit operations and the way the article evolves in response to an edit, for their scoring. For instance, *reverts* (undoing an edit) and *restores* (changing back to an earlier version) are treated as direct indications of respectively distrust and trust in most of these work; depending on the reputation of the initiator editor, the reputation of the recipient editor is identified.

Text stability is exploited by other authors, [2,3,21] with the intuition that an edit or text that remains longer as part of the article has been approved implicitly by other editors compared to an edit that is undone very soon, hence implying some notion of trust. Having access to visit log data, the number of visits is another metric used as the notion of the quality or impact of a contribution in [26].

Other features that are used to establish some notion of trust are the reputation of the editors of previous versions [10,34], interactions in other contexts such as admin elections and barnstars [24], and finally the degree of intervention of admins in monitoring and improving articles [17].

Our work is also related to coordination and conflict modeling in Wikipedia. For instance, Kittur et al. [20] uses several article-level metrics such as the number of authors, the number of versions, the number of anonymous edits, etc. to train a model for predicting the number of controversial tags assigned to an article. Similarly, using these tags as the ground truth, Vuong et al. [31] built a model to assign a controversy score to articles assuming a mutual reinforcing relationship between controversy score of articles and their contributors.

Identifying attitudes in other domains. Considerable work has been done in *sentiment analysis* in many different domains, ranging from product reviews [25] to forum discussions and news [8,29]. Sentiment analysis aims at classifying statements about a topic, event or product as either positive or negative, but does not directly extract relationships between people. However, it can be used to classify people based on the similarity of their stated opinions to supporting or opposing camps [29].

On a more related level, there is also work on extracting agreement or disagreement relationships in conversational meetings or discussion forums [11, 14, 15]. For meetings in particular, these relationships are learned from a wide range of features such as the number of words in the utterance of the first speaker, the number of common n-grams in their utterances [11,15], and the previous history of agreement or disagreement between the two speakers [11]. For forums, the attitudes of participants are inferred from the sentences that contain second person pronouns and a sentiment word [14]. Depending on the sentiment expressed in these sentences, a positive or negative attitude is assigned between the corresponding participants of each sentence.

Social Network Analysis. Our work is also related to link prediction or inference in both unsigned and signed networks. Gilbert et al. [12] used seven different categories of features and a linear regression model to learn the strength of links between users in Facebook. The model was trained on a dataset with 2000 links collected from responses of 35 participants about the strength of their relationships with random members of their friends list. In another work [32], a notion of link strength between users was established based on the pairwise similarity of their profiles and their interaction histories and was tested in both Facebook and LinkedIn.

Finally, the link prediction was studied in a more recent work [23] in the context of two well-known theories from social psychology, namely *balance*, and *status*, and was tested on three different domains: Slashdot, Epinions, and Wikipedia admin election.

3. TERMINOLOGY

In this section, we define some of the key technical terms that we use later in the paper.

Definition 1. *We say an interaction happens between two editors e_1 and e_2 if they both edit the same article and their edits are related.*

Two edits may be considered related under multiple circumstances. For example, two edits may be considered related if they are applied to the same or near-by sections of an article; also one would expect related edits to happen within some temporal proximity with not many other edits falling between them. This is on the basis that an edit may be triggered by or may rectify to improve an earlier edit, and with a large gap between two edits, this hypothesis may not hold. For our purpose of inferring interactions between editors, the following gives a more clear and workable definition of relatedness.

Definition 2. *Two edits are related if they are both applied to the same article and the distance between the edits in terms of the number of intermediate versions is less than a threshold.*

The distance between edits is defined in terms of the number of versions the article goes through between the edits, instead of elapsed time, to account for the variance that exists in activity rate of different articles. To set the threshold for related edits, we look at the maximum distance of all pairs of edits that modify the same *section*[1] of the same Wikipedia article. Intuitively, two versions editing the same section are more likely to be the result of a collaboration interaction and be related than two edits on different parts of the same article or different articles.

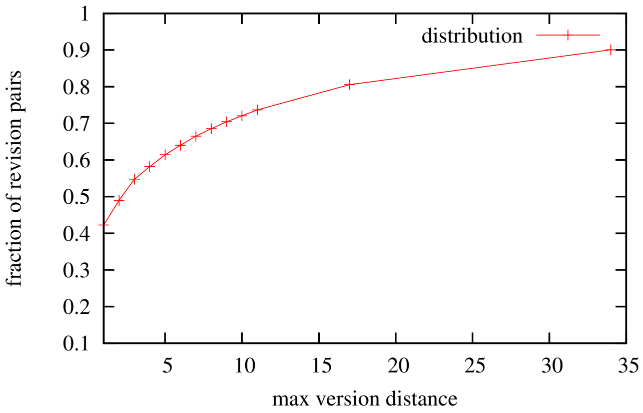


Figure 2: Distance distribution of revision pairs editing the same section

Figure 2 shows the cumulative distribution of the pairs of revisions editing the same section in terms of the maximum distance (i.e. number of versions) in between. The data for this plot comes from a random sample of 100 Wikipedia articles. As one can see, these revisions in about 40% of times are happened in two consecutive versions, while 70% are applied within 9 or less versions, and finally over 90% of them are at most 34 versions apart. Thus, we set the threshold for edits to be related at 34.

Definition 3. *Collaboration profile is an ordered summary of a set of statistics about the individual activities of each of the e_1 and e_2 editors on editing Wikipedia articles, along with their pairwise interactions on the set of their coedited articles.*

For each collaboration profile, cp_{e_1,e_2}, a sign (positive or negative) can be assigned denoting the supporting or opposing attitude of editor e_1 toward editor e_2. This attitude can be potentially different from the attitude of e_2 toward e_1 (cp_{e_2,e_1}). Hence, in the profile of cp_{e_1,e_2}, we refer to e_1, and e_2 as the *source* and the *target* editors respectively.

Definition 4. *A collaboration network is a directed, signed graph $G = (V, E)$ associated with a Wikipedia article a, where V is the set of all editors who contributed in creating at least one revision for a, and $E \subset V \times V \times W$ is the set of weighted edges connecting editors with non empty collaboration profiles.*

A directed edge from e_1 towards e_2 with weight w represents an existence of a cp_{e_1,e_2}, and w is a number that can be positive or negative depending on the type of the collaboration profile. In a *binary* signed collaboration network, $w \in \{-1, +1\}$, while in a *weighted* signed collaboration network, w can be any real number, often normalized within $[-1, +1]$.

[1] Every article in each of its versions can be broken down into shorter units, called sections, where each section discusses the article from a different aspect.

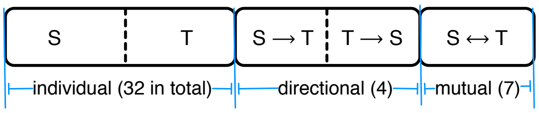


Figure 3: Three groups of features in collaboration profile representing the attitude of editor S (Source) towards T (Target)

4. INFERRING ATTITUDES

With a large number of revisions associated to an editor, and an even larger number of words, expressions or clues that may indicate some form of sentiment in those revisions, it becomes a challenge to infer the attitude of one editor towards another. In this section, we describe our approach for inferring attitudes using a set of features and statistics obtained from revision history of edited pages. These statistics are used in a form of a collaboration profile, which we learn to classify them as positive or negative denoting the corresponding attitude of one editor towards another. In the following two sections, we first describe how we build these profiles, and then how we learn to classify them.

4.1 Building collaboration profiles

The features used in our collaboration profiles are grouped into three parts, as shown in Figure 3. As the names suggest, individual features are derived from each editor individually based on his/her revisions, and represent the general behavior of that editor, while the directional and mutual features represent the behavior of an editor with respect to another editor, describing how this editor (referred to as source editor) interacts or collaborates with the other editor (referred to as target editor). Hence, individual features are assigned to each editor, the directional group features are assigned to an ordered pairs of editors, whereas the mutual features are assigned to an unordered pairs of editors. In the following, we describe each of these groups of features in more detail.

Individual features. To describe the general activity level of an editor, our individual features include for each editor the number of articles edited, total number of revisions, and the average contribution size over articles edited, where the contribution size of an editor in an article is the ratio of revisions made by the editor to all revisions made to the article. As a base to see the type of articles an editor edits, the average concentration ratio of all articles edited by the editor is also kept (i.e. concentration ratio of an article is defined as the ratio of unique editors to all revisions of that article). The intuition here is that a high concentration ratio might cause bias and information censoring compared to a low concentration ratio where the contribution is shared among all editors more evenly.

To characterize the general writing tone of an editor, our features include the number of agreement terms and the number of disagreement terms both in comments of revisions made by the editor, selected from the most frequent words including uni-grams, bi-grams and tri-grams that appear in the comment lines of a set of manually-tagged agreement and disagreement edits. Prominent examples of terms indicating agreement are “add”, “fix”, “spellcheck”, “copyedit”, “clarify”, and “move”. On the other hand, terms such as “uncited”, “fact”, “is not”, “bias”, “claim”, “revert”, and “see talk page” are indicators of disagreement. Similar statistics are also extracted from revisions immediately after an editor’s revisions, which are likely to contain responses.

Other features we consider are based on revert and restore actions, for example whether or not other editors like an editor’s revisions. Specifically, we keep for each editor the number of reverted revisions, the number of restored revisions and also the number of revisions made by others but reverted by that editor. We also keep track of the average Δ size of an editor’s revisions and the average Δ size of revisions immediately after his revisions. The Δ size of a revision is the net change in length of the article, measured in characters between consecutive revisions , hence it can be negative if an edit simply removes some content.

In order to capture *a priori* tendency of an editor to get into conflicts, we also compute the average conflict score of all articles which the editor has been involved with, where the conflict score of an article is simply the fraction of *conflicting* interactions in that article.

Definition 5. *An interaction is* conflicting *if any one of these conditions are met:*

1. *the revisions are consecutive and the edit length of the later revision is negative,*
2. *the revisions are consecutive and the later revision uses more negative terms than positive terms in its comment line, or*
3. *the revisions are related and the later revision reverts the earlier revision.*

Finally, we also compute the average time it takes an editor to respond as well as the average time before the editor gets a response. Intuitively, these can help gauge if an editor engages in so-called “revision wars”, characterized by rapid fire of disagreement responses.

Directional features. As for directional features, we have separate statistics for each pair of editors (one in each direction). For these features, we use the ratio of co-edited articles to all articles edited by one editor, and the ratio of revisions by one editor in co-edited articles to the edits of that editor across all articles he edited.

Mutual features. Finally, for the class of mutual statistics we treat the pair undirected by considering the following features: the number of co-edited articles; the number of interactions; ratio of conflicts over all interactions; the average number of versions between related revisions in corresponding interactions; and the fraction of interactions corresponding of consecutive revisions; Also as a base for comparison, we compute the average concentration ratio, and conflict score of all co-revised articles for the pair of editors.

4.2 Classifying collaboration profiles

Given the profiles of editors and their collaborations, our goal is to classify each collaboration into one of *agreement* or *disagreement.* In the absence of labeled data, one needs to resort to heuristics to infer labels for collaborations. For instance, Maniu et al. [24] use features such as the number of deleted, inserted, and replaced words, and whether an editor has given barn-star award to another and label each feature intuitively as a sign of a positive or negative relationship. Then, the final sign of the relationship of a pair of editors is determined based on the sign of the majority

Table 1: Statistics of election dataset

number of elections	3713
number of unique editors	9541
positive votes	130193
negative votes	36239
repeated votes	5601
conflict votes	1420

class. The authors in [5] develop a content-based method, by building a topic model of edits. In their method, the relationship of a pair of editors, for instance, editing the same paragraph takes a value in the range [-1,1] depending on whether one editor changes the topic distribution of the paragraph towards the changes made by the other editor of that paragraph or not. This approach again relies on a heuristic which is limited to interactions that can change the topic distribution of an article; the method also has not been evaluated.

Our work takes a more systematic approach by leveraging the strong relationship that exists between the way Wikipedia editors collaborate in editing pages, and the way they later vote in admin elections; the intuition here is that an editor who casts, for example, a negative vote to a candidate is more likely to have a negative than a positive interaction with the candidate. In fact, this dataset is used in the work of [24] with votes being a deciding feature in the sign of relationships between two editors. However, they could only use this feature for pairs who participate in elections, and the number of those pairs is much smaller than the number of pairs who interact.

In this paper, we use the election data to learn the weight of features that contribute to positive or negative collaborations. More specifically, we use the election data and tag a limited set of interactions as positive or negative; a classifier is built on this labeled data, which can then be used to predict the sign of collaboration profiles for other editors who may or may not appear in the elections dataset. Our results show that such a classifier can be built with a high accuracy which is evident of the influence of collaboration interactions on votes.

5. PREDICTING ADMIN ELECTIONS

In this section, we describe how we can learn collaboration profiles from admin elections and how these profiles can be useful in predicting votes. In our discussions, *administrators* (shorten as admins) refer to a set of editors in Wikipedia with certain, higher (than ordinary editor) privileges who are chosen by regular elections. In particular, in each election, an editor becomes a candidate for promotion into an admin editor and other editors can cast supporting, opposing or neutral votes towards that candidate. The information about admin elections is available from Wikipedia and is also used in some recent work [22, 23]. Our dataset (as explained next) is more up-to-date and includes more elections.

The election data is available in the usual Wikipedia dump in the form of special articles called "Requests for Adminship" (RFA). We collected and parsed all these RFA articles from a recent Wikipedia dump (dated April 5, 2011), resulting in a dataset that covered 3713 elections (compared to 2794 elections extracted up to January 2008 in [23]). More statistics about this dataset are shown in Table 1.

Table 2: Number of votes in extracted and mapped election data

	extracted data	mapped data
total	166432	89652
positive	130193	75168
negative	36239	14484

5.1 Learning profiles and predicting votes

With collaboration profiles as our feature vectors and votes in admin elections as the corresponding labels, we train a classifier to learn the relationship between profiles and vote signs. More specifically, for each candidate c and voter v, we build their profile $cp_{v,c}$ and their respective features (as discussed in section 4).

There are a few caveats in building collaboration profiles to predict elections. First, since we want to predict the sign of the votes before they are cast, a time constraint in building collaboration profiles is to use only the information that is available prior to a vote. Second, a candidate and a voter can appear in multiple elections possibly at different times and the vote of the candidate can change from one election to the next. In cases where v casts a vote for c only once in the entire election dataset, all revisions up to the time of casting vote seem to be relevant and are used for building collaboration profiles. Similarly, in cases where v casts the same vote for c multiple times (referred to as *repeated votes*), all revision history up to the time of vote is considered. However, for multiple conflicting votes (referred to as *conflict votes*), as the vote of v casted for c change over time, we consider only the revision history from the time of the most recent vote v casts for c.

5.1.1 *Mapping collaboration profiles to votes*

Our collaboration network is built over pairs of editors who collaborate in revising articles (or more precisely, have related edits), and these pairs may or may not appear in the election data. Our experiments on predicting elections only considers pairs who have collaboration profiles and also appear in the election data (these pairs are referred to as *mapped data* in Table 2).

Table 2 shows the number of votes that are extracted from the election data, the number of votes that could be mapped to our feature vectors and their break-down to positive and negative votes. The ratio of positive votes to all votes is 78% and 83% in extracted and mapped datasets respectively.

5.2 Results

Table 3 shows the performance results of our approach on predicting votes using a 10-fold cross validation experiment on full and balanced mapped dataset. The balanced-dataset is obtained from the full dataset by randomly sub-sampling positive votes until the number of positive and negative votes are the same. For these results, we tested four classifiers, namely Random Forest, J48, SMO and Logistic using Weka [2] machine learning tool.

As we can see in Table 3, the Random Forest classifier achieves the highest accuracy among the studied classifiers in both datasets. In fact, Random Forest classifiers have a good performance in general and also on imbalanced datasets, as shown in some previous work [19], due to their bagging

[2] www.cs.waikato.ac.nz/ml/weka

Table 3: Results of predicting votes on full and balanced datasets.

Model	F-Acc.	F-AUC	B-Acc.	B-AUC
Random Forest	**0.869**	**0.877**	**0.781**	**0.857**
J48	0.842	0.706	0.695	0.707
SMO	0.838	0.5	0.579	0.579
Logistic	0.837	0.626	0.591	0.628
All positive	0.838	0.5	0.5	0.5
Relative-edit	0.82	0.5	0.499	0.5
talk-positive	0.537	0.570	0.583	0.583

Table 4: 15 most important features of the vote classifier

of candidate's agreement terms
of candidate's edits
candidate's avg. contrib. size
candidates's avg. time being responded
of candidate's disagreement terms
avg. Δ size of revisions after candidate's
avg. Δ size of candidate's revisions
candidate's avg. response time
of candidate's edits reverted
avg Δ size of voter's revisions
of candidate's edited articles
of reverts made by the candidate
avg concent. ratio in articles edited by candidate
avg concent. ratio in articles edited by voter
avg version distance of interactions

and internal feature selection methods. Hence, for this classifier, we applied an additional tuning and feature selection method by following the approach proposed in [27]. In particular, for ranking and selecting features, we used the Gini importance metric of the classifier, and removed 7 features with the lowest importance score. These features were 1) number of co-revised articles, 2) number of interactions, 3) fraction of interactions corresponding to consecutive revisions, 4) average conflict score of co-edited articles, 5) average concentration ratio of co-edited articles, 6) ratio of revisions by candidate in co-edited articles to all of his revisions, and 7) same as 6 but for voter. After selecting features, we tuned the two parameters of the classifier which led us to choose 70 trees and 15 random features at each branch.

This additional tuning and feature selection resulted into 86.9% and 78.1% accuracy, an about 1% improvement (which translates to about 1000 more correct predictions in our data) and 8% over the default setting of Random Forest in Weka on full and balanced datasets respectively. Table 4 shows the top 15 features which are ranked by importance according to the Random Forest classifier.

As is shown, the features that are ranked on top are mostly individual features, and are that of the candidate; this is consistent with our intuition as individual activities of the candidate is more influential on the outcome of votes than the characteristics of the voter. Top 15 features include from interaction features only the average version distance of interactions which is representative of the strength of interactions of the two editors modeled by how fast the two editors responded or collaborated with each other in revising the articles.

5.2.1 *Comparison with other methods*

We also compared our method with three simple baselines: all-positive, relative-edit and talk-positive: all-positive classifies all votes as positive; relative-edit classifies a vote as negative if the number of previous edits of candidate and the voter is in the same range (having the difference of less than 10), and positive otherwise; finally, talk-positive classifies a vote as positive if the candidate and the voter have any previous communication (writing to each other's user pages).

The relative-edit and talk-positive baselines are from [22], where authors speculated that voters in admin elections tend to do a relative assessment of candidate through implicit comparison to their own merits such as comparing their number of edits with the candidate's (as a sort of the level of activity). The authors also found that having previous communication increases the probability of positive votes.

Comparing all these methods on the full dataset, we can see that our best results using the Random Forest classifier shows about 3% and 37% improvement over the strong all-positive baseline in terms of respectively accuracy and area under ROC curve (which is a measure commonly used for imbalanced datasets [19]). On the balanced dataset, the superiority of our method is more pronounced: 25% higher than the best baseline. We observed that relative-edit has almost the same performance as all-positive. This is because while the probability of negative votes increases when the relative edit distance of candidate and voter is small, the number of pairs with such condition is very low, which makes this baseline to act like the all-positive. On the other hand, as about half of all pairs had a prior communication, the talk-positive baseline cannot be very effective in explaining most of the votes. This shows that while there are important factors that increase the probability of giving positive votes, it is unlikely that a single factor can explain different reasons of different positive and negative votes, and thereby a combination of several factors should be used to learn these votes.

We cannot compare our results against those reported by Leskovec et al. [23] as their method is not applicable when the voter or the candidate does not appear in past elections. Even for pairs of candidates and voters that appear in past elections, their method has a tuning parameter called minimum *embededness* (defined as the number of common neighbors) and both their accuracy and the votes they can predict vary with the values of this parameter, whereas in our work, the interaction window-size is a parameter affecting the votes that can be predicted. These differences make a fair comparison very difficult in general. That said, our method on balanced dataset shows 78.1% accuracy, compared to 80.1% reported by Leskovec et al., which supports an argument that the analysis of revision history (as we do) can be as valuable as the analysis of the social network of voters (as they do) for predicting admin elections.

5.2.2 *Effect of interaction window size*

A parameter used in our method is the size of the window where two interactions are considered related. This parameter is set by default to 34 (as discussed in section 3). However, a question is how our method performs as we vary the window size. With a change in window size, clearly the number of votes that can be mapped will change. Our results show that by increasing the window size up to 40, the number of votes that can be mapped also increases; after this point, an increase in window size has a very small effect. As expected, by considering larger window sizes, we allow editor pairs to have an interaction in farther distances; this in turn increases their chances of having at least one interaction, but after some point, a larger window size does not increase the chance of finding any new interactions, which is consistent with our definition of related edits.

Also, studying the trend of accuracy over different window sizes, we found out that the classifier is quite stable and by increasing the window size from 2 up to 60, there is less than 3% drop in accuracy. Hence, we can conclude that increasing the window size allows more votes to be predicted for the expense of small loss in accuracy, and also a bit longer time for extracting features and training the classifier.

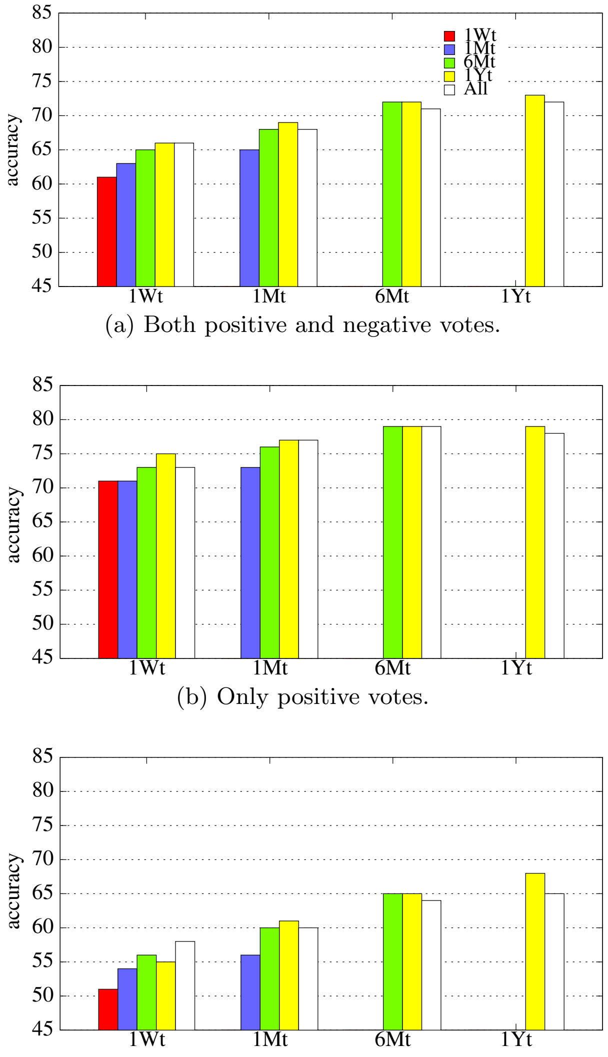
5.2.3 *Effect of history length*

In previous experiments, we built the profiles of our editors and their collaborations using all interactions before the casting of a vote. A question that arises is if all edits and interactions are relevant when one wants to predict votes. For example, an interaction that happens much before an election may not carry much weight. In this section, we want to limit the length of history that can be used to construct profiles, and to find out how the performance of our method is affected. It should be noted that limiting the relevant history length also changes the set of predictable votes since we need at least one co-edited article in the history to be able to build our profiles.

Figure 4 shows the change in accuracy on the balanced dataset as we vary the history length. A problem here is that as the history length changes, the set of votes that can be predicted also changes and this makes a comparison difficult. To address this problem and to keep the votes the same, as the history length is increased, even though more interactions are used in building our profiles, we limit our prediction only to pairs of editors who show interactions in our shorter windows. Each cluster of bars in our figures shows the prediction result with the votes kept the same but the length of the history varied. For instance, the cluster on the far left shows the results when the votes are restricted for pairs of editors who show interactions within one week before the vote. The label *all* denotes the case where the entire history is used to build our profiles.

Looking at figure 4, our first finding is that even on balanced dataset, the positive accuracy for all history lengths and vote sets is much higher than the negative accuracy. For instance, there is more than 10% difference between positive and negative accuracy of votes of 1 week when they are learned using all interactions. This implies that in general it is harder to explain negative votes based on previous history of interactions, and other hidden factors such as current state of casted votes, and the response of the candidate to asked questions during the election may play a role in the votes that are cast.

The other interesting finding is that using a longer history improves all accuracy metrics for votes of all set of votes. However, this increasing trend usually stops or even gets reversed after one year period. This suggests that one year is a suitable length for capturing most of the collaboration attitudes of Wikipedia editors and any information about the collaboration of editors beyond this time is not vital for predicting votes.



(a) Both positive and negative votes.

(b) Only positive votes.

(c) Only negative votes.

Figure 4: Trend of accuracy over all, positive and negative votes across different time periods. The x-axis represents the vote sets and the bars in each cluster is the prediction result for different history lengths.

Also, we see that whenever longer history improves the overall accuracy, negative votes benefit more than positive votes and the overall difference of bars in each cluster for negative accuracy is higher than positive accuracy. For instance, considering the votes of the first two clusters where we have the most change in metrics, the accuracy of the negative class in overall relatively improves 15% and 8% in each cluster respectively, while the positive accuracy for the same clusters has 5% and 4% relative improvement by using larger history.

This difference in the effect of history length on positive and negative votes might suggest that if the voter casts negative votes based on previous revision history of collaborations, he might remember and refer to negative events from a long time before the vote time. In contrast, positive votes can be explained even by the recent history of collaborations.

6. IDENTIFYING CONTROVERSIES

In this section, we validate our method by showing its usefulness in the task of identifying controversial articles. In particular, for this task, we first build a collaboration profile for each pair of editors who interact in editing the article (based on our discussion in Section 4). We then classify all the collected collaboration profiles associated with each article a by using our vote predictor classifier. The vote classifier trained on admin election data assigns a positive or negative label to each collaboration profile, which represents the sign of the edge connecting the corresponding editors in the collaboration network of each page. Furthermore, for building collaboration profiles and connecting editors in our collaboration networks, we imposed an additional constraint of having at least two edited versions for each editor. This additional constraint aims at removing occasional, non-active editors and giving a more reasonable number of possible collaborating pairs.

Our validation task here is identifying controversial articles, based on the observation that these articles are tagged as controversial by Wikipedia editors themselves. Note that unlike some previous work [20,31], where the number of controversial tags assigned during the history of the page was used as the evaluation method, we consider modeling the problem as a simple binary classification task (whether the page is controversial or not). Controversial tags are tags such as {controversial}, {dispute}, {disputed-section}, etc. that can be added to the text of a version by Wikipedia editors. We avoid using these tags as the ground truth since they can have several problems such as forgetting to remove the tag after the controversy resolves or to add a tag long after the start of controversy as pointed out in [30]. Hence, the number of these tags is not necessarily representative of the controversy degree of pages.

Automatically identifying these articles can benefit both editors and article readers by warning them about the disputed state of the article and how they should interpret the content or manage the collaboration process. Also, due to highly disputed nature of these pages, we expect to have different structure of agreement and disagreement relationships between editors in the corresponding collaboration networks compared to other pages which provides a suitable test set for our method of inferring attitudes.

6.1 Overview

For our purpose, we aim to show that our approach for building collaboration networks leads to different structure between networks of controversial articles and other articles. Previously, Brandes et al [6] showed the structural difference of controversial articles using a metric called *bipolarity*. Bipolarity is a graph-based metric that measures how much a graph is likely to decompose into two opposing groups, where most of disagreement edges will be between the two groups rather than within them. Hence, one approach would be to extract bipolarity from the collaboration networks we build and compare its values over controversial articles and non-controversial articles. However, as bipolarity is defined only for negative edge networks, and was also shown to not provide enough discrimination between controversial and non-controversial pages [28], we extracted some other features from our collaboration networks instead of focusing on a single metric.

More specifically, we set up a classification task where for each controversial page c, and similarly for each non-controversial page n, we extract a set of features f_1, f_2,...,f_k obtained from the corresponding collaboration networks of each page. Then, we train a classifier to learn to distinguish controversial and non-controversial articles using these features where each page (network) is an instance in our task. We show more accuracy is obtained for this classification task using the features extracted from our built collaboration networks compared to other methods for building collaboration networks.

6.2 Selecting pages

We carefully selected 240 articles for each group of controversial, and non-controversial by the following procedure: We selected controversial articles by randomly selecting pages from all the 15 categories in the list of controversial articles maintained in Wikipedia [3]. When choosing articles, we chose the new title of pages in case of redirected pages. From the 240 selected articles, only 122 articles had controversial tags in their revision history.

For selecting non-controversial pages, we randomly chose 100 pages from the featured article category, and 150 pages from the other quality groups. For each of these pages, we also check that they will not be among controversial articles list. This is because many of the articles in the list of controversial articles later become non-controversial, and even improve to featured articles due to several factors such as limiting the editors access. We avoid choosing such pages for our non-controversial set, and consider only pages without any controversy in any part of their revision histories (i.e. not only in the current state of the page).

6.3 Features extracted from collaboration networks

We extract the following 30 features in total from the graph of collaboration network associated with each controversial or non-controversial article in our test set:

- total number of non-isolated nodes (isolated nodes are nodes not connected to any other node in the graph)
- total, positive, and negative number of edges
- average of total, positive, and negative degrees of nodes
- the percentage of nodes having a in-degree of higher than 90% of maximum in-degree (one feature for each positive and negative in-degree), and similarly for out-degree
- the percentage of nodes having an in-degree of less than 110% of minimum in-degree (one feature for each positive and negative in-degree), and similarly for out-degree
- the percentage of nodes having an in-degree of in the range of 10% lower and 10% higher than the average in-degree (one feature for each positive and negative in-degree), and similarly for out-degree
- the percentage of nodes with higher positive than negative in-degree, and similarly for out-degree
- total number of triads

[3] http://en.wikipedia.org/wiki/List_of_controversial_articles

- the relative number of each of the 8 triad types

Triads are directed sub-graphs of size 3, which have been used as important metrics in many recent work such as [13, 18, 23]. We considered 8 different triad types in our work depending on how many negative edges exists (0, 1, 2, or 3) and whether the edges in the triad form a cycle or not.

6.4 Comparison with other methods

The first method that we compared our method with is a simple, and well-known method for building collaboration networks [6], in which authorship is considered at the word level. Based on this notion, whenever editor e_1 deletes some words originally inserted by e_2 in the text of article a, an edge with a negative weight proportional to the number of words deleted will be created from e_1 to e_2. Also, whenever e_1 restores a version created by e_2 to an earlier version created by e_3 (a possibly different editor than e_1), a single unit positive edge will be created from e_1 to e_3, and a single unit negative edge will be created from e_1 to e_2 as e_1 undo the work of e_2 and implicitly agrees with the work of e_3. Note that while we specifically compare our method with this method of building collaboration networks, using delete, and revert actions as disagreements between editors is a common method in other work on Wikipedia [2, 16, 17, 34].

We also compare our method with three baselines: a) Rand50 refers to a method that randomly assigns positive, or negative sign to each of the edges in the network built by our method. b) Rand83 is another baseline that assigns the sign of edges randomly similar to baseline1, but with a probability of 83% positive, and 17% negative, which are the ratio of positive and negative votes in the full mapped admin election. c) NE-count is a method that only uses the number of nodes, and the number of edges in our built networks as features, and does not use any information about the structure of the links.

Finally, we compare our method with a previously proposed method for classifying controversial articles which uses meta features such as the number of versions, the number of reverts, etc. obtained from the revision history of each article [28]. We refer to this method as *Meta* (and exclude two disagreement related features from this method to only focus on meta features).

6.5 Results

Table 5 shows the accuracy of the classifier in detecting controversial articles using each of the methods. The results are based on 10-fold cross validation using Logistic classifier. For our method, collaboration networks built using the vote classifier trained on balanced dataset showed slightly better results than the one trained on the full data-set and hence we report our method based on this better result. Also, the results for Rand50 and Rand83 methods are the average results over 20 runs with different random seeds.

First, as we can see the NE-count baseline has the lowest accuracy among all methods showing that the number of authors and their interactions are not alone enough to distinguish controversial articles from other articles and the actual network structure matters.

Second, comparing our method with other methods for building collaboration networks, we see a substantial improvement. Specifically, our method has more than 20% higher accuracy than the DRR method. This suggests that considering interactions at smaller unit level (word-level authorship) and also relying only on basic edit operations cannot capture different collaboration relationships between editors, and more extensive global features across all co-edited articles is needed for inferring these relationships. Also, comparing with the two Rand methods, where we have the same network structure and features as our method, and just the sign of edges is different across these methods strongly supports the importance of inferring attitudes and the effectiveness of our method for doing so.

Table 5: Results of identifying controversial pages.The same 30 features were used in DRR, Rand50, Rand83 and our method.

Method	Acc
NE-count	56.70%
DRR	64.31%
Rand50	68.67%
Rand83	71.31%
Meta	75.20%
our method	**84.58%**
our method + Meta	**89.12%**

Finally, comparing with the Meta method, we see that while general features about the revision history provide a good discrimination between the two studied classes of articles, they cannot eliminate the important role of the structural properties of the collaboration network of editors. In fact, by taking advantage of these two complementary views (structural and meta features), we are able to boost the performance of both methods and achieve a very promising results of 89.12% for this task.

7. CONCLUSION

In this paper, we showed that revision history of Wikipedia articles contains valuable information that can be utilized in different tasks related to Wikipedia. In particular, we showed that there is a strong correlation between the previous collaboration history of editors and how they vote for each other in admin elections. This new perspective about admin elections in Wikipedia allowed us to not only be able to predict votes with a high accuracy, but to use this dataset as a training environment for inferring attitudes of editors. Besides, studying the relationship of votes and previous history of collaboration of editors showed interesting differences between positive and negative votes, where we found that positive votes usually can be explained by even the recent history of interactions, while negative votes might be associated with negative interactions from long time before the time of the vote.

As an application of inferring attitudes, we tested our method on identifying controversial articles based on the structural properties of the signed collaboration networks of articles, built using our attitude classifier. Comparing with a previous attempt on modeling editors relationships, and also with a method based on meta data of revision history, our promising results suggested that the structural properties of collaboration networks and modeling the attitudes of editors beyond the simple edit operations is indeed crucial for understanding the collaboration nature of Wikipedia articles.

Acknowledgements. This work was supported in part by grants from NSERC.

8. REFERENCES

[1] Spectral analysis of signed graphs for clustering, prediction and visualization. In *SIAM'10*, pages 559–570, 2010.

[2] B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to wikipedia content. In *WikiSym '08*, pages 1–12, 2008.

[3] B. T. Adler and L. de Alfaro. A content-driven reputation system for the wikipedia. In *WWW '07*, pages 261–270, 2007.

[4] Alexa.com. http://www.alexa.com/topsites. Last visited on Oct 28, 2011.

[5] P. Bogdanov, N. D. Larusso, and A. Singh. Towards community discovery in signed collaborative interaction networks. In *ICDMW'10 workshop*, pages 288–295, 2010.

[6] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in wikipedia. In *WWW '09*, pages 731–740, 2009.

[7] K. Chatterjee, L. de Alfaro, and I. Pye. Robust content-driven reputation. In *AISec '08*, pages 33–42, 2008.

[8] Y. Choi, Y. Jung, and S.-H. Myaeng. Identifying controversial issues and their sub-topics in news articles. In *PAISI*, volume 6122, pages 140–153, 2010.

[9] N. Clark. Trust Me! Wikipedia's Credibility Among College Students. *International Journal of Instructional Media*, 38(1):27–36, 2011.

[10] G. M. Druck and A. G. McCallum. Learning to predict the quality of contributions to wikipedia. In *WikiAI08*, pages 7–12, 2008.

[11] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. Identifying agreement and disagreement in conversational speech: use of bayesian networks to model pragmatic dependencies. In *ACL'04*, 2004.

[12] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *CHI'09*, pages 211–220, 2009.

[13] M. Granovetter. The strength of weak ties: A network theory revisited. In *Sociological Theory*, pages 105–130, 1982.

[14] A. Hassan, V. Qazvinian, and D. Radev. What's with the attitude?: identifying sentences with attitude in online discussions. In *Empirical Methods in Natural Language Processing*, pages 1245–1255, 2010.

[15] D. Hillard, M. Ostendorf, and E. Shriberg. Detection of agreement vs. disagreement in meetings: training with unlabeled data. In *HLT-NAACL 200*, pages 34–36, 2003.

[16] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. Measuring article quality in wikipedia: models and evaluation. In *CIKM '07t*, pages 243–252, 2007.

[17] S. Javanmardi, C. Lopes, and P. Baldi. Modeling user reputation in wikis. *Statistical Analysis and Data Mining*, 3(2):126–139, 2010.

[18] Katherine and Faust. A puzzle concerning triads in social networks: Graph constraints and the triad census. *Social Networks*, 32(3):221 – 233, 2010.

[19] T. M. Khoshgoftaar, M. Golawala, and J. V. Hulse. *An Empirical Study of Learning from Imbalanced Data Using Random Forest*, volume 2, pages 310–317. 2007.

[20] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi. He says, she says: conflict and coordination in wikipedia. In *CHI '07*, pages 453–462, 2007.

[21] T. R. Korsgaard and C. D. Jensen. Reengineering the wikipedia for reputation. *Electronic Notes in Theoretical Computer Science*, 244:81 – 94, 2009.

[22] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Governance in Social Media: A case study of the Wikipedia promotion process. In *ICWSM'10*, 2010.

[23] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW'10*, pages 641–650, 2010.

[24] S. Maniu, B. Cautis, and T. Abdessalem. Building a signed network from interactions in wikipedia. In *DBSocial'11 workshop*, 2011.

[25] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[26] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in wikipedia. In *GROUP'07*, GROUP '07, 2007.

[27] D. M. Reif, A. A. Motsinger, and B. A. McKinney. Feature selection using a random forests classifier for the integrated analysis of multiple data types. In *CIBCB*, pages 1–8, 2006.

[28] H. Sepehri Rad and D. Barbosa. Towards identifying arguments in wikipedia pages. In *WWW'11*, pages 117–118, 2011.

[29] S. Somasundaran and J. Wiebe. Recognizing stances in online debates. In *ACL'09*, pages 226–234, 2009.

[30] R. Sumi, T. Yasseri, A. Rung, A. Kornai, and J. KertÃl'sz. Edit wars in wikipedia. *Technology*, pages 724–727, 2011.

[31] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, and H. W. Lauw. On ranking controversies in wikipedia: models and evaluation. In *WSDM'08*, pages 171–182, 2008.

[32] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *WWW'10*, pages 981–990, 2010.

[33] B. Yang, W. Cheung, and J. Liu. Community mining from signed social networks. *IEEE Transactions on Knowledge and Data Engineering*, 19:1333–1348, 2007.

[34] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness. Computing trust from revision history. In *PST '06*, page 1, 2006.