Towards Identifying Arguments in Wikipedia Pages

Hoda Sepehri Rad
Department of Computing Science
University of Alberta,Canada
sepehrir@ualberta.ca

Denilson Barbosa
Department of Computing Science
University of Alberta, Canada
denilson@ualberta.ca

ABSTRACT

Wikipedia is one of the most widely used repositories of human knowledge today, contributed mostly by a few hundred thousand regular editors. In this open environment, inevitably, differences of opinion arise among editors of the same article. Especially for polemical topics such as religion and politics, difference of opinions among editors may lead to intense "edit wars" in which editors compete to have their opinions and points of view accepted. While such disputes can compromise the reliability of the article (or at least portions of it), they are recorded in the edit history of the articles. We posit that exposing such disputes to the reader, and pointing to the portions of the text where they manifest most prominently can be beneficial in helping concerned readers in understanding such topics. In this paper, we discuss our initial efforts towards the problem of automatic evaluation of extracting controversial points in Wikipedia pages.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: web-based services

General Terms

Algorithms, Evaluation, Classification, Experimentation

1. INTRODUCTION

Currently, Wikipedia handles controversy by allowing editors to (manually) tag *entire articles* as controversial, thus informing both editors and readers about the disputed reliability of such articles. To help in this process, there have been recently some attempts to automatically detect controversial articles as well as to rank them based on their degree of controversy [1, 2, 3]. However, such approaches work at the whole article level which is very coarse to help users in understanding the actual controversies. We posit that knowing which sections or paragraphs of an article concentrate the most controversial content would be more helpful to the readers.

We aim identifying the most controversial points in a Wikipedia article to allow both readers and editors judge the article and its existing opinions better. In other words, let p be a Wikipedia article with n different topics described at some point in the history of p. Then, our problem is to find and summarize the top-k most controversial *arguments* in p's history. An argument is concise sub-set of the article (e.g., a section or a paragraph); we say an argument is controversial if its corresponding edit graph can be partitioned into groups of users such that there is high agreement between users within the same partition but little agreement between users across partitions. For instance, at the time of writing, two controversial arguments in Wikipedia's article about abortion are the sections "the breast-cancer hypothesis", and "personhood".

Besides modelling disagreements and finding controversial arguments, one challenge in our work is arriving at a reliable way of measuring how controversial an argument is. Since Wikipedia keeps only a list of highly controversial articles (without information regarding the controversial arguments themselves), we lack a human-prepared gold standard to rely on. Also, assessing the extracted arguments based on a human judgment experiment by tracking edits and discussions can be very time-consuming and, in the end, the results would be subjective. In this paper, we study whether objective metrics for identifying controversial articles at the whole page level can be used for assessing the degree of controversy within individual arguments.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

2. DESIDERATA

We start by defining desiderata for a measure of controversy $C(\cdot)$ to be suitable to our task. First, $C(\cdot)$ must be consistent with Wikipedia's classification of controversial articles; we refer to this requirement as the *discrimination* criterion. Second, and most importantly, $C(\cdot)$ must exhibit a monotonic decrease in value as more and more arguments are removed from an article. That is, if p is an article and p' is obtained by removing an argument from p, then we we must have that $C(p') \leq C(p)$. We refer to this second requirement as the *monotonicity* criterion.

The intuitive justification for insisting on discriminating measures of controversy is that we want the results to be consistent (to the extent possible) with what editors and readers are used to. Similarly, the justification for requiring the controversy measures to be monotonic is that removing parts of an article can only remove some of the disagreement present in that article (and never introduce new disagreement).

3. POSSIBLE CONTROVERSY METRICS

We have studied two controversy measures with respect to the criteria above. The first measure, called bipolarity [1], computes how close the *edit graph* of an article is to a perfect bipartite graph; intuitively, this measure indicates whether the editors of such an article can be divided into two oppos-

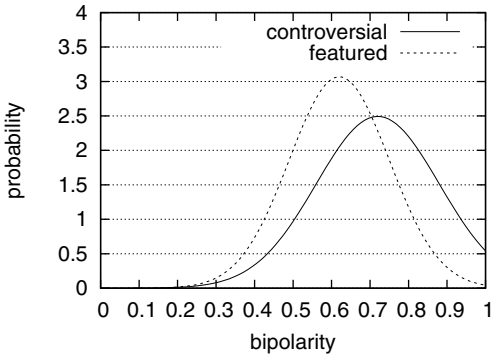

Figure 1: Probability distribution functions of bipolarity metric for two sets of pages

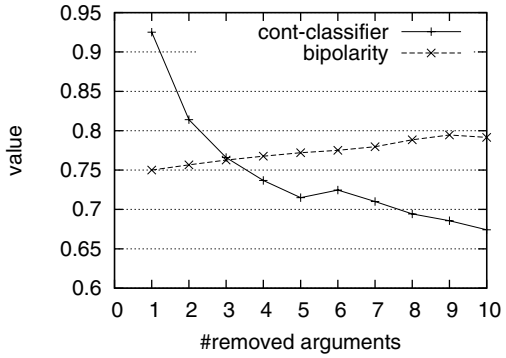

Figure 2: Monotonicity test for the studied metrics

ing camps. The authors show that on the average, bipolarity is higher for controversial articles compared to "featured" articles in Wikipedia[1].

The second measure we consider relies on a classifier trained on a set of controversial and featured articles. We train the classifier using controversial articles as positive examples, and featured articles as negative examples. We use the degree of membership to the controversial class that is assigned by the classifier to each instance as the actual measure of controversy. We refer to this metric as cont-classfier.

The actual features used by the classifier are as follows (AVG, STD, MAX, and SUM mean average, standard deviation, maximum and sum, respectively): absolute number of (1) versions, (2) minor versions, (3) anonymous editors, (4) unique editors, (5) versions per editor; relative frequency of (6) versions by anonymous editors; AVG, STD and MAX for the (7) number of versions per user; AVG, STD, MAX, and SUM of (8) disagreement scores in the edit graph (computed as in [1]); and AVG, STD of (9) length of edits in each revision.

4. RESULTS

Discrimination. Figure 1 shows the distribution of the bipolarity metric for 60 (randomly selected) controversial and 60 (randomly selected) featured articles[2]. As can be seen from this figure, although the average of the distribution function of controversial articles is higher than featured articles, the variances are high and the two distributions overlap substantially. Using larger sample of articles, consisting of the 200 controversial and 200 featured (including the 120 used to produce Figure 1), we were able to achieve 60% accuracy for the discrimination power of bipolarity.

Using the same set of 400 articles as in the experiment above, we trained our cont-classifier using the Random Forest algorithm with k=5 and I=70. The accuracy of the resulting classifier in a 10-fold cross validation experiment was 85%. Thus, we conclude that the cont-classifier has a much higher discriminative power.

Monotonicity. In order to test the monotonicity property of the measures above, we proceed as follows. For each article, we sort the sections by decreasing number of edits; then, we repeatedly remove each section (in order) and record the relative drop in value for the measure.

Figure 2 shows the results of testing the monotonicity condition using a set of 60 random controversial pages. Observe that bipolarity fluctuates with the removal of sections, increasing slightly after the removal of 10 sections by almost 5%. On the other hand, the value of cont-classifier decreases monotonically, although not linearly: there is a more pronounced decrease up to removing the 5th argument, at which point some minor fluctuations are noticed, followed by a slower rate of decrease. In any case, the decreasing trend is evident.

5. CONCLUSION

In our work, we are investigating ways of finding controversial issues within a single Wikipedia article, unlike previous works that focus on controversies at the whole page level. We discussed two necessary properties for an evaluation metric and studied whether state-of-the-art controversy metrics for whole pages satisfy these criteria.

Our results show that the cont-classifier measure satisfied both desiderata, and seems promising as a general indication of argument controversy. Our next step is determining its sensibility in reflecting the difference of different argument selector methods and disagreement models, and to find out how much it is consistent with known controversial issues.

Acknowledgements. This work was supported in part by a grant from the NSERC Business Intelligence Network.

6. REFERENCES

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

[2] 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.

[3] 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.

[1] An article is considered featured if it meets all quality standards of Wikipedia for accuracy, neutrality and coverage.

[2] These were the exact same articles, from the same version of Wikipedia used in [1].