Characterization and prediction of Wikipedia edit wars

Róbert Sumi, Taha Yasseri, András Rung, András Kornai, János Kertész
Institute of Physics
Budapest University of Technology and Economics
H-1111 Budapest Budafoki u 8
{rsumi,yasseri,runga,kertesz}@phy.bme.hu, kornai@math.bme.hu

ABSTRACT

We present a new, efficient method for automatically detecting conflict cases and test it on five different language Wikipedias. We discuss how the number of edits, reverts, the length of discussions deviate in such pages from those following the general workflow.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces**]: Group and Organization Interfaces—*Collaborative computing, Computer-supported cooperative work, Web-based interaction*; K.4.3 [**Computers and Society**]: Organizational Impacts—*Computer-supported collaborative work*

General Terms

Algorithms, Measurement, Languages

Keywords

Wikipedia, Collaboration, Conflict, Classification

1. INTRODUCTION

Wikipedia (WP) is among the largest and highest impact web 2.0 sites based on the collaboration of millions of contributors (called *editors*). With its freely accessible full documentation it is subject of scientific research on topics from artificial intelligence like taxonomy questions [1] to sociology of popularity [2]. As in other walks of life, the course of collaboration does not always run smoothly and, especially in the case of contentious issues and highly politicized subjects, we find serious conflicts, called in WP parlance *edit wars*. Our interest is with the statistical detection of such conflicts against the background of the overwhelming majority of pages which show peaceful development and constructive conflict resolution.

For the human viewer of page histories it is evident that an article such as `Liancourt Rocks`[1], discussing a group of small islets claimed by both Korea and Japan, or the article on `Homosexuality` were the subject of major edit wars. Yet articles with a similar number or relative proportion of edits such as `Benjamin Franklin` or `Pumpkin` were, equally evidently to the human reader, developed peacefully. Conflicts in WP were studied already both on the article and on the user level. Kittur [3] et al. computed article controversy from different page metrics (number of reverts[2], number of revisions etc.), Vuong et al. [4] counted the number of deleted words between users and used their "Mutual Reinforcement Principle" to measure how controversial a given article is. Both teams counted how many times the 'controversial' tags [5] appeared in the history of an article, and used this as ground truth. While this is an excellent test in one direction (certainly recognition of controversiality by the participants is as good as the same recognition coming from an outsider), it is too narrow, as there can be quite significant wars that the participants are unaware of or at least do not tag, as, e.g., in the articles on `Gdańsk` or `Euthanasia`.

As in most pattern recognition tasks such as speech or character recognition, we take human judgment to be the gold standard or *truth* against which machine performance is to be judged. Put this way, the characterization task is simply a binary classification problem. Our approach is to construct a numerical measure M of controversiality that we use to rank WP pages and to characterize those subject to edit wars with the goal of detecting impending conflicts. An additional benefit of this approach is that once we have a reliable measure of controversiality we can automatically select high- and low-controversiality populations and investigate how the two differ. While the examples are from the English WP, in order to create a robust, language- and culture-independent measure we analyzed not just the English, but also the Czech, Hungarian, Persian and Spanish WPs by the same methods.

We assume the reader to be aware of the structure of WP (article pages, discussion pages, user pages, talk pages) and some of the internal policy guidelines such as `Wikipedia: Neutral point of view`, but we do not assume the reader to have deep familiarity with WP lore.

[1]Throughout this paper, references in `typewriter font` are to WP.

[2]On Wikipedia, reverting means undoing the effects of one or more edits to a version that existed sometime previously.

2. DETECTING EDIT WARS

Our detection method is entirely based on reversion, but the raw revert statistics do not yield a clear cutoff-point we could use to distinguish controversial from non-controversial articles. There are several confluent criteria that mark pages as edit wars, including (i) overt notices requesting cleanup or deletion, (ii) lengthy talk pages, (iii) (repeated) freezing of the page, (iv) involvement of senior editors in WP's internal arbitration processes, and (v) bans on some of the editors from working on the page or on WP as a whole. Yet, the same type of argument as above makes clear that none of these criteria are sufficiently strong in isolation to unambiguously distinguish war from peace in WP. Rather than building a complex but arbitrary formula that includes several of these factors, our goal is to base the decision on very few parameters – ideally, just one.

Let be $\ldots, i-1, i, i+1, \ldots, j-1, j, j+1, \ldots$ consecutive revisions in the history of an article. If the text of revision j coincides with the text of revision $i-1$, we considered the revert between the editor of revision j and i respectively. We are interested in disputes where editors have different opinions about the topic, and do not reach consensus easily. Let us denote by N_i the total number of edits in the given article of that user who edited the revision i. We characterize reverts by pairs (N_i^d, N_j^r), where r denotes the editor who makes the revert, and d refers to the reverted editor (self-reverts are excluded). Fig. 1 represents the revert map of the non-controversial `Benjamin Franklin` and the highly controversial `Israel and the apartheid analogy` articles. Each mark corresponds to one or more reverts. The coordinates of the marks are the total number of edits of the reverter (N^r) and the reverted editor (N^d). Clearly, the disputed article contains more reverts between editors having large edit numbers than the uncontroversial article.

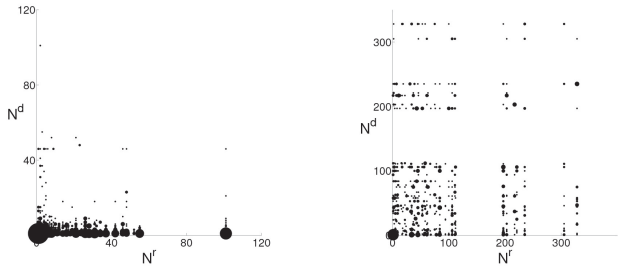

Figure 1: Revert maps of the articles `Benjamin Franklin` and `Israel and the apartheid analogy`. N^r and N^d are the total number of edits of the reverter and reverted editor respectively. The size of the mark is proportional to the number of reverts between them.

The revert maps already distinguish disputed and non-disputed articles, and we can improve the results by considering only *mutual* reverts. This causes little change in disputed articles (compare the right panels of Fig. 1 to that of Fig. 2), but has great impact on non-disputed articles (compare left panels).

Based on the rank (total edit number within an article) of editors, two main revert types can be distinguished: when one or both of the editors have few edits to their credit

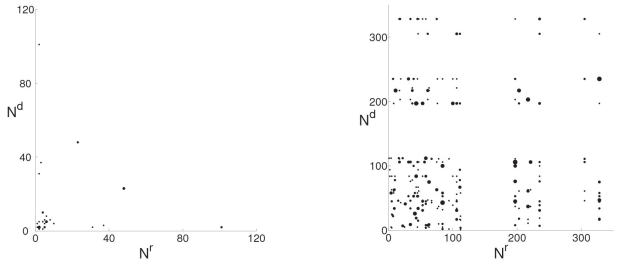

Figure 2: Maps of mutual reverts in the same articles as in Fig. 1.

Table 1: Precision of controversiality detection based on number of edits #e, reverts #r, mutual reverts mr, talk page length TP, raw M_r, M_i, article tag count TC, and M.

WP	#e	#r	#mr	TP	M_r	M_i	TC	M
`cs`	14	18	26	28	25	27	27	28
`en`	27	29	29	30	26	28	30	28
`hu`	4	27	28	26	23	29	24	30
`fa`	24	28	26	28	29	29	25	28
`es`	23	26	29	29	27	28	28	29
%av	61	85	92	94	87	94	89	95

(these are typically reverts of vandalisms since vandals do not get a chance to achieve a large edit number as they get banned by experienced users) and when both editors are experienced. In order to express this distinction numerically, we use the *lesser* of the coordinates N^d and N^r, so that the total count includes vandalism-related reverts as well, but with a much smaller weight. Thus we define our raw measure of controversiality as

$$M_r = \sum_{(N_i^d, N_j^r)} min(N_i^d, N_j^r).$$

Once we developed our first autodetection algorithm based on M_r, we iteratively refined the controversial and the non-controversial seeds on multiple languages by manually checking pages scoring very high or very low. In this process, we improved M_r in two ways: first, by multiplying with the number of editors E, resulting in M_i (the larger the armies, the larger the war) and second, by censuring the topmost mutually reverting editors (eliminating cases with conflicts between two persons only). Our final measure of controversiality M is thus defined by

$$M = E \sum_{(N_i^d, N_j^r) < max} min(N_i^d, N_j^r).$$

We have checked this measure for five different languages and concluded that its overall performance is superior to other measures, including the presence of the 'controversial' WP tags both in terms of precision and recall.

3. PREDICTING EDIT WARS

Roughly speaking, controversiality rears its head at around $M = 50$, meaning that only one page in a hundred becomes even a candidate for war (less than 30k out of over 3m articles in the English WP). Less than .5% of pages shows significant signs of war ($M > 200$) which suggests that a good method of predicting impending wars is to monitor crossing the $M = 100$ threshold. Our work in this area is still ongoing.

4. ACKNOWLEDGMENT

This work was supported by the EU's 7th Framework Program's ET-Open within ICTeCollective project no. 238597. Special thanks to Santo Fortunato for discussions and his help with data at early stages of this work.

5. REFERENCES

[1] S. P. Ponzetto and Strube M. Knowledge derived from Wikipedia for computing semantic relatedness *Journal of Artificial Intelligence Research* 30:181-212, 2007.

[2] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer and A. Vespignani Characterizing and modeling the dynamics of online popularity *Physical Review Letters* 105, article no. 158701, 2010.

[3] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: conflict and coordination in Wikipedia. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462, New York, NY, USA, 2007. ACM.

[4] Ba-Quy Vuong, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, and Hady Wirawan Lauw. On ranking controversies in wikipedia: models and evaluation. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 171–182, New York, NY, USA, 2008. ACM.

[5] Wikipedia `http://en.wikipedia.org/wiki/Wikipedia:Template_messages/Disputes`