Identifying Controversies in Wikipedia using Support Vector Machines

Ba-Quy Vuong
baquy@cs.wisc.edu
Computer Sciences Department - University of Wisconsin at Madison

Abstract

Wikipedia is a very large and successful Web 2.0 example. As the number of Wikipedia articles and contributors grows at a very fast pace, there are also increasing disputes occurring among the contributors. As a result of disputes, many articles in Wikipedia are controversial. In this project, I propose a supervised learning model using Support Vector Machines (SVMs) to identify controversial articles in Wikipedia. The idea is to represent each article by a bag-of-word feature vector. Each value in this vector is a raw count of a word type appearing in the article. Experiments on real articles from Wikipedia show that the proposed approach can effectively identify controversial articles.

Introduction and Motivation

Using open source Web editing software (e.g. wiki), online community users can now edit, review and publish articles collaboratively. Among the large and more successful wiki sites is Wikipedia (Voss July 2005), the online encyclopedia which covers 16.6 million articles (both English and non-English), 9.5 million users and 200 languages (Wikipedia a). As Wikipedia is growing very fast in both number and size, there is also a higher likelihood for disputes to occur among contributors. Disputes often happen in articles with controversial content, in which contributors have different or even opposite opinions. For example, “Iraq War” is one of the most controversial articles in Wikipedia. It attracts a lot of disputes among contributors because they have different standing points about the war. Some people support it while some others strongly oppose to it. They also have difficulties in agreeing on different facts of the war. Additionally, disputes can be caused by “defensive” contributors who always argue to defend their ideas even when they are incorrect. As a result of disputes, many articles in Wikipedia are controversial.

In this project, I aim to identify controversial articles (*controversies* for short), which is important due to the following two reasons. First, controversies appearing in Wikipedia articles are often a good reflection or documentation of the real world. Finding controversies in Wikipedia can therefore help the general public and scholars to understand the corresponding real world controversies better. Second, It allows moderators and contributors to quickly identify highly controversial articles, thereby improving the effectiveness of the dispute resolution process by reducing the amount of effort searching for such articles.

However, determining controversies in Wikipedia is a great challenge. First, there is a huge number of articles, which makes it impossible to manually look at each article and identify controversies. Second, the articles cover a wide range of topics that require a lot of background knowledge for identifying controversies. Third, articles are growing very fast and that makes any results ever obtained be soon outdated.

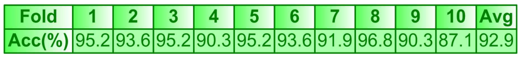
There have been several approaches related to determining controversies in Wikipedia. Wikipedia currently lets users to assign controversial tags (Wikipedia b) to articles to signal its controversies. Some other works, including (Lim et al. 2006; Hu et al. 2007b; 2007a; Vuong et al. 2008; Adler and de Alfaro 2007; Kittur et al. 2007) focus on developing unsupervised models to rank articles' qualities and controversies in Wikipedia. However, these approaches are either inefficient or achieving very low accuracy.

In this project, I propose an automatic approach for identifying controversies in Wikipedia. In particular, I represent each article as a bag-of-word vector in which each value is the raw count of a word type. I then apply a supervised learning model based on SVMs to learn models for detecting controversies. These learned models are subsequently used to classify unlabeled articles. Experiments on real articles from Wikipdia show a great promise of this solution in detecting controversies.

Methodology

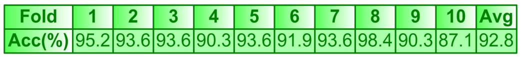
Articles: In Wikipedia, each article consists of a sequence of revisions. Each revision is created whenever a contributor makes some changes and saves them. The article's content that we see in Wikipedia is just its latest revision. Conceptually, two different revisions of an article can have different content as they may possibly contain different words or the same words but in different orders. Thus, two different revisions of a single article may have different controversy characteristics. In my approach, by “article” I refer to a particular article revision at a certain time.

Bag-of-word representation: In my approach, each article in the data set is first segmented into a bag of words. The data set's vocabulary is created by merging all word types



Fold	1	2	3	4	5	6	7	8	9	10	Avg
Acc(%)	95.2	93.6	95.2	90.3	95.2	93.6	91.9	96.8	90.3	87.1	92.9

Figure 1: Accuracies of the first model.



Fold	1	2	3	4	5	6	7	8	9	10	Avg
Acc(%)	95.2	93.6	93.6	90.3	93.6	91.9	93.6	98.4	90.3	87.1	92.8

Figure 2: Accuracies of the second model.

appearing in all articles. After the vocabulary is created, each article is represented by a bag-of-word vector, whose size is equal to the vocabulary size. For each vector, the i^{th} value corresponds to the i^{th} word type in the vocabulary, and that value captures the number of occurrences of the corresponding word type in the article. Note that I just keep the raw word counts and do not perform any normalization.

Additional features: In Wikipedia, controversial articles often have longer content (i.e. the number of words in the latest revision) or longer histories (i.e. the number of revisions) than non-controversial ones. This is because controversial articles often attract more attention from community users. As a result, they add/remove content from these articles more often and that causes articles to have longer histories. However, the editing activities due to controversy do not happen forever. After some "peak" editing periods, an article often stabiles (because the contributors have found consensus on controversial content, or the moderator has prevented editing due to editing wars (Wikipedia c)). These stabilized articles often have long content. To capture these properties, I design two additional features that represent article content length and article history length. In the current implementation, I only consider article content length and leave the history length for a future work.

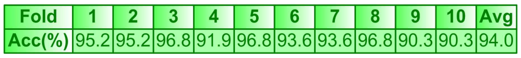
Data sets: To collect data sets for training, I retrieve a set of controversial articles and a set of non-controversial ones from Wikipedia. The process of collecting data and performing preprocessing is discussed in detailed in the next section.

Training and Classifying: After collecting labeled articles and generate feature vectors for them, I train the model using SVMLight (Joachims) software. The resulting model is saved to a file and is subsequently used for classifying test data and unlabeled data.

Empirical Evaluation

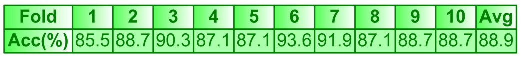
Data collection: Wikipedia has a special page listing the most controversial articles in each category (Wikipedia d). To download controversial articles, I crawled for articles from the Politics/Economics category and only kept their latest revisions. To download non-controversial articles, I crawled for other articles from the same category. To ensure these articles are truly non-controversial, I only kept the articles with reasonable histories (more than 100 revisions) and had not been tagged as controversial. Again, I only retained the latest revisions. There were 130 controversial articles and 525 non-controversial articles in total.

Data preprocessing: After articles were downloaded, each article is an xml document which contains both the article's wiki text and a lot "dirty" data (such as xml tags, wiki formatting tags, punctuations, stop words, etc). To ensure these dirty data do not affect the learned models, I performed several cleaning steps. First, I removed all the xml tags, wiki formatting tags and their associated text. Next, I removed all characters which are neither letters nor digits. After that, I segmented each document into a bag of words and removed all stop words from this bag.



Fold	1	2	3	4	5	6	7	8	9	10	Avg
Acc(%)	95.2	95.2	96.8	91.9	96.8	93.6	93.6	96.8	90.3	90.3	94.0

Figure 3: Accuracies of the third model.



Fold	1	2	3	4	5	6	7	8	9	10	Avg
Acc(%)	85.5	88.7	90.3	87.1	87.1	93.6	91.9	87.1	88.7	88.7	88.9

Figure 4: Accuracies of the fourth model.

Models: Next, I applied SVMs to learn four models and compared their performance. The first model uses just bag-of-word vector representation and a linear kernel. The second model uses bag-of-word vector representation but with the polynomial kernel $(sa^Tb+c)^d$ where $s=1, c=1, d=2$. The third model is the same as the first model except it adds article length as an extra feature. Finally, the fourth model extends the second model by adding the extra article length feature. To evaluate the performance of each model, I performed a ten-fold cross validation.

Runtime: Training the first and the second models on the collected data sets was very fast. Each fold under these models took less than one minute. However, training the third and the fourth was much longer. Each fold took roughly about 30 minutes. This fact indicates that when the article length is added as a feature, the training algorithm converges much more slowly.

Accuracies: Tables 1,2,3, and 4 show the accuracy of each fold and the average accuracy of the each model. It is clear that all the models perform very well in detecting controversial and non-controversial articles. The average accuracies are much higher than the baseline accuracy of 0.8[1]. The addition of the article length into the third and fourth models has different effect. In the third model, it helps improve the accuracy while in the fourth model, it reduces the accuracy. Among the first two models, the first model also has a slightly better performance. That implies the use of a polynomial kernel does not help to make the two article sets more separable.

Most influential features: Looking at the parameter vector (w) of each fold, I observe that the component with the greatest magnitude in nine folds corresponds to the word "party". That means "party" plays an important role in classifying controversial articles. The remaining word is "education". This result is reasonable because the category of the data sets is Politics/Economics. Thus, words which are highly related to this category play important roles.

Conclusions

Automatically identifying controversies in Wikipedia is a challenging problem due to many factors. In this project, I propose a supervised learning model using Support Vector Machines to identify controversial articles in Wikipedia. In this approach, each article is represented by bag-of-word features and some additional features. Experimental results show that this is a very promising direction to proceed.

[1] In a baseline model, we just classify all articles as non-controversial and the accuracy obtained is $526/(526+130) = 0.8$

References

Adler, B. T., and de Alfaro, L. 2007. A content-driven reputation system for the wikipedia. In *WWW*, 261–270.

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

Hu, M.; Lim, E.-P.; Sun, A.; Lauw, H. W.; and Vuong, B.-Q. 2007b. On improving wikipedia search using article quality. In *WIDM*, 145–152.

Joachims, T. `http://svmlight.joachims.org/`.

Kittur, A.; Suh, B.; Pendleton, B. A.; and Chi, E. H. 2007. He says, she says: conflict and coordination in wikipedia. In *CHI*, 453–462.

Lim, E.-P.; Vuong, B.-Q.; Lauw, H. W.; and Sun, A. 2006. Measuring qualities of articles contributed by online communities. In *Web Intelligence*, 81–87.

Voss, J. July 2005. Measuring wikipedia. In *International Conference of the International Society for Scientometrics and Informatics*.

Vuong, B.-Q.; Lim, E.-P.; Sun, A.; Le, M.-T.; and Lauw, H. W. 2008. On ranking controversies in wikipedia: models and evaluation. In *WSDM*, 171–182.

Wikipedia. `http://en.wikipedia.org/wiki/Size_of_Wikipedia`.

Wikipedia. `http://en.wikipedia.org/wiki/Template_talk:Controversial`.

Wikipedia. `http://en.wikipedia.org/wiki/Revert_wars`.

Wikipedia. `http://en.wikipedia.org/wiki/List_of_controversial_articles`.