Using Machine Learning to Augment Collaborative Filtering of Community Discussions

(Extended Abstract)

Michael Brennan
Dept. of Computer Science
Drexel University
3175 JFK Blvd Room 140
Philadelphia, PA 19104
mb553@cs.drexel.edu

Stacey Wrazien
Dept. of Computer Science
Drexel University
3175 JFK Blvd Room 140
Philadelphia, PA 19104
saw42@cs.drexel.edu

Rachel Greenstadt
Dept. of Computer Science
Drexel University
3175 JFK Blvd Room 140
Philadelphia, PA 19104
greenie@cs.drexel.edu

ABSTRACT

Collaborative filtering systems have been developed to manage information overload in online communities. In these systems, users rank content provided by other users on the validity or usefulness within their particular context. Slashdot is an example of such a community where peers rate each others' comments based on their relevance to the post. This work extracts a wide variety of features from the Slashdot metadata and posts' linguistic contents to identify features that can predict the community rating. We find that author reputation, use of pronouns, and author sentiment are salient. We achieve 76% accuracy at predicting the community's rating of the post as good, neutral, or bad.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*multiagent systems*

General Terms

Measurement, Experimentation, Human Factors

Keywords

collaborative filtering, machine learning, trust, reputation

1. INTRODUCTION

Today's online communities have developed a variety of community-based filtering and rating mechanisms to help maintain quality and manageability. In general, we can refer to these systems as collaborative filtering systems. The goal is that "good" content will rise to prominence and "bad" content will fade into obscurity. These filtering mechanisms are not well-understood and have some known weaknesses. For example, they depend on the presence of a large crowd to rate content, but such a crowd may not be present. Additionally, the community's decisions determine which voices will reach a large audience and which will be silenced, but it is not known if these decisions represent "the wisdom of crowds [5]" or a "censoring mob [3]." Our approach uses statistical machine learning as a way to objectively gain insight into the workings of these filtering mechanisms. By extracting features that replicate their workings, we can better understand collaborative filtering, improve the way the community uses the ratings of their members, and design agents that augment community decision-making.

In this paper, we study the Slashdot (slashdot.org) community, identifying a combination of features which allow us to extract comments that the community will rate as good with high (82%) accuracy. Furthermore, we can segment comments into good, neutral, and bad categories with 76% accuracy. We found that author reputation and contextual features were the most salient, however, we also discovered many salient linguistic features, which, when used alone can extract good comments with 57% to 63% accuracy, depending on the inclusion of humorous posts.

Cite as: Using Machine Learning to Augment Collaborative Filtering of Community Discussions (Extended Abstract), Michael Brennan, Stacey Wrazien, Rachel Greenstadt, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 1569-1570

2. METHODOLOGY

We chose to mine our data from Slashdot (slashdot.org), a technology news site and online community. Readers of the site submit articles which are reviewed by a team of editors, who select the best ones to post as the news items for that day. The community then discusses the articles and issues posted through a comment system. Each news post has its own comment series. Slashdot has implemented a collaborative filtering system for users to rank the comments on how relevant they are to the article and to other users on a scale from -1 to 5, with 5 signifying the comments most worth reading. This system has been studied in the past [1, 2]. Comments that receive a very low score are typically hidden, while comments with a higher score are highlighted, allowing the user to easily reach quality commentary. In addition to the numerical rating posts can also be given a rating description such as "Insightful" if it is good and "Offtopic" if it is bad, among others.

The features we used to classify Slashdot comments are divided into two groups: linguistic features and contextual and author reputation features. The linguistic set represents features related to the words, their meanings, and the structure of the text. Most of the linguistic features were extracted from the comments using the Linguistic Inquiry and Word Count (LIWC) software, a text analysis database designed by psychologists to study various emotional, cognitive, and

structural components of verbal and written speech [4]. The contextual and author reputation features are based upon information such as when it was posted or how much discussion it generated, or information about the author such as what his or her recent comment ratings have been. A full list of features can be found on the web[1].

We evaluated the ability of our feature set to predict the community rating of comments made on Slashdot news stories on the dates of Saturday, February 14th and Monday, February 16th 2009. All classification was performed using a SVM Classifier that used a Gaussian radial basis function kernel. The features were all discretized into four bins before being used for classification (except LIWC_sentiment which already had three discrete values). We took samples from a data set of 528 comments or 1173 depending on whether we divided the data set into two classes or three. This variation is due to the need to keep the class distribution equal and changing the score range for each class affected the maximum number of comments that could be selected. Accuracy measurements were obtained by running each experiment five times, generating a new data set and feature space for each iteration. Classification was performed using the WEKA toolkit[2] and the LIBSVM library[3].

3. RESULTS

While the Slashdot rating system allows for comments to be rated from -1 to 5, we found that attempting to classify a comment as belonging to a specific score class is not very useful - there is too much noise involved and the benefits of classifying something as a 4 instead of a 5 is negligible towards improving the quality of discourse. So we looked at two different methods of categorizing the comments: extracting the good comments and ignoring the rest, and dividing the comments into "good," "neutral," and "bad" categories.

3.1 Extracting Good and Bad Comments

We considered a comment to be of the highest quality if it had a rating equal to or higher than three. Using our extended feature set we were able to determine whether or not a comment was rated in this highest set by the community with 82% accuracy. This demonstrates the ability of a machine learning system to perform the most important task for a collaborative filtering system meant to enhance the level of discourse about a topic: highlighting the elements of the discussion which are most relevant and worthwhile.

But only extracting the good comments is not necessarily enough for an effective agent meant to augment collaborative filtering systems. We do not necessarily want to penalize comments that would be deemed by the community to be simply "average." We examined the ability of our classifier to specifically segment the comments between "bad", "neutral", and "good" posts. A "bad" post is one with a score of -1 or 0, a "neutral" post has a score of 1, and a "good" post has a score greater than or equal to two. We were able to classify the comments with an overall accuracy of 76%, significantly higher than random-chance classification of 33%.

3.2 Comparing Linguistic Features to Contextual and Author Reputation Features

[1]http://psal.cs.drexel.edu/files/Slashdot_Features.pdf
[2]http://www.cs.waikato.ac.nz/ml/weka/
[3]http://www.csie.ntu.edu.tw/ cjlin/libsvm/

The most salient features in our set are contextual features such as subCommentCount and author-reputation features like posterRecentScore. We found that when we looked at linguistic features alone they were not as effective as the contextual and reputation based features but were still quite salient in determining the community rating of a comment, especially features like first-person pronouns and comment word length. This is especially true when comments that are classified as "funny" are left out. Humorous comments often have a very different linguistic makeup when compared to "informative" or "interesting" comments, leading to linguistic features being less effective when classifying them.

In the case of extracting "good" comments with a score greater than or equal to three, linguistic features alone yielded an accuracy of 55%. If we removed "funny" comments, however, that accuracy rose to an average of 63%. The task of segmenting the comments between "good", "neutral", and "bad" yielded an accuracy of 42%. Once again if we removed the "funny" comments we saw an increase to 46%.

4. CONCLUSION AND FUTURE WORK

This work demonstrates that machine learning can be a valuable tool for gaining an objective understanding of how values are embedded in technologies, how communities develop reputations and norms, and how socio-technical communities can combine human and machine computation. The work we have done thus far with the Slashdot data set has shown that author past performance (reputation) is a good proxy for future results. However, the linguistic feature results suggest that there are interesting and unexpected features to be found that can provide insight into the workings of these community filtering mechanisms. Even in an irreverent community like Slashdot, "I-statements" are indicators of good content and civility matters. Our results show that the work of moderators can be amplified by machine learning techniques as we are able to achieve 76% accuracy (precision and recall) in replicating their assessments. This accuracy is made possible by the structure and metadata of online communities. The 42% accuracy achieved using linguistic features alone shows that finding interesting commentary automatically is an interesting and likely achievable goal for natural language processing.

5. REFERENCES

[1] C. Lampe and P. Resnick. Slash(dot) and burn: Distributed moderation in a large online conversation space. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2004.

[2] C.A. Lampe. *Ratings Use in an Online Discussion System: The Slashdot Case*. PhD thesis, University of Michigan, 2006.

[3] Annalee Newitz. The censoring mob : How social media destroy freedom of expression - and why that might be a good thing. In *Hacking at Random (HAR)*, 2009.

[4] James W. Pennebaker, Martha E. Francis, and Roger J Booth. Linguistic inquiry and word count - liwc2007. www.liwc.net, 2007.

[5] James Surowiecki. *The wisdom of crowds : why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.* Doubleday, 2004.