Learning to Extract Quality Discourse in Online Communities

Michael Brennan, Stacey Wrazien and **Rachel Greenstadt**
Dept. of Computer Science
Drexel University
3175 JFK Blvd Room 140
Philadelphia, PA 19104
tel 215.895.2920
fax 215.895.0545
{mb553,saw42,greenie}@cs.drexel.edu

Abstract

Collaborative filtering systems have been developed to manage information overload and improve discourse in online communities. In such systems, users rank content provided by other users on the validity or usefulness within their particular context. 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 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" or a "censoring mob." Our approach uses statistical machine learning to predict community ratings. By extracting features that replicate the community's verdict, 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. Slashdot is an example of such a community where peers will 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 predicting community ratings as good, neutral, or bad.

Introduction

We are witnessing a transition from a world in which gatekeepers and editors filter content before it is published, to a world full of vast amounts of user-generated content in which information filtering is done after publication. Online communities have developed a variety of community-based filtering and rating mechanisms to help maintain quality and manageability. James Surowiecki notes that several conditions must be present (including diversity and independence) for a crowd to be wise (Surowiecki 2004). It is an open question whether these collaborative filtering mechanisms represent "the wisdom of crowds (Surowiecki 2004)" or "the censoring mob (Newitz 2009)."

In collaborative filtering systems, users rank content provided by other users on the validity or usefulness within their particular context. The goal is that "good" content will rise to prominence and "bad" content will fade into obscurity. These mechanisms are heterogenous in their design, ranging from the simple (up or down vote) to the byzantine. The principles underlying these mechanisms are not well understood and they have several known weaknesses. First, the content must have a large enough initial audience to produce a rating. Even on popular sites, the critical mass necessary to produce ratings dissipates quickly, so that comments on articles not added quickly are destined for obscurity. Second, these mechanisms are often gameable. Work by Annalee Newitz showed how `digg` could be manipulated to get arbitrary content on the front page (Newitz 2007). Third, despite the use of these mechanisms, information overload persists. Last, these mechanisms often enshrine existing voices and overlook content from new or marginalized voices.

Our approach uses statistical machine learning techniques to predict outcomes automatically and objectively gain insight into the workings of these filtering mechanisms. We show that though understanding the content of a piece of writing is difficult, determining whether an online community will find it interesting is more tractable. Our ultimate goal is not to produce automated systems to replace collaborative filtering, but to design mechanisms by which individuals, communities, and intelligent software agents can collaborate to explain and improve social, bottom-up filtering, and expand the range of the possible in terms of the values these systems can reflect and the communities it can serve.

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[1]. 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. These features provide insight into what types of posts garner high versus low ratings.

[1] Throughout this paper "accuracy" refers to the precision averaged across all classes. Additionally, recall in all cases was within one percent of the precision.

Related Work

There are many well-studied issues surrounding online communities that rely on peer rating systems to determine the relevance of the content. Information overload can easily occur in online communities when it becomes difficult for users to filter the relevant and interesting content. This can encourage users to leave the community (Farzan, DiMicco, and Brownholtz 2009). In addition, considerable time can pass before the fair and poor comments are identified in communities like Slashdot (Lampe and Resnick 2004). This research illustrates why it is important to have a valid and accurate rating system.

A well-studied superset of our topic is recommender systems, which use collaborative filtering and/or learning to rate and recommend arbitrary products or content (Herlocker et al. 2004; Massa and Aversari 2007; Terveen and Hill 2001). This research has noted the difficulty in identifying the combinations of measures to use in a comparative evaluation (Herlocker et al. 2004). There has been considerable interest and work by Paul Resnick on designing recommender systems that are not vulnerable to manipulation (Resnick and Sami 2008; 2009). Another large body of related work deals with assigning reputation to individuals rather than their work (Dingledine, Freedman, and Molnar 2000; Josang, Ismail, and Boyd 2007; Levien 2004; Xiong and Liu 2004).

Other related work has taken an ethnographic approach to studying the phenomenon of online spaces (Finin et al. 2007; Harper 2009; Surowiecki 2004), some of it looking directly at the Slashdot community (Lampe 2006). In "Social Network Sites: Definition, History, Scholarship (danah boyd and Ellison 2007)," danah boyd and Nicole Ellison tackle the question of defining social networks and online communities.

The work differs from related research in that it deals with the task of learning how a community selects quality discourse and applies it to automatically improve and facilitate discussions. The goal of this research is to use machine learning and eventually an agent to augment online discussions based on an understanding of how the community works, instead of simply using a peer-reviewed method.

The Slashdot Community

Slashdot (slashdot.org) is 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 through a comment system. Each news post has its own comment series. Due to the large number of comments each article receives, Slashdot has implemented a trust metric for users to rank the comments on how relevant they are to the article and to other users. The Slashdot community is a useful community to start with because their community rating system is particularly rich in metadata. The system uses a scale from -1 to 5, with 5 signifying the comments most worth reading. Comments that receive a very low score are typically hidden, while comments with a higher score are highlighted. This is beneficial because it prevents a user from sorting through an abundance of useless data in order to access relevant commentary. In addition to the numerical rating posts can also be given a rating description such as "Insightful" or "Informative" if it is good and "Offtopic" or 'Flamebait" if it is bad, among many others.

There are many other nuances to the Slashdot system such as moderator points and meta-moderation that are not discussed in this paper but complete details about how the Slashdot rating system works can be found in their FAQ[2].

Slashdot was chosen because the richness of the rating scheme helps make the site a valuable testbed for an agent intended to augment a collaborative filtering system. Our goal, however, is not to augment Slashdot's system alone. We believe that this research can open the door to augmenting systems with more obscure or less robust rating systems.

A sample news story[3] and highly rated (5) comment[4] based on it follows:

> Subject: Rabbit Ears To Stage a Comeback Thanks To DTV
> Post: Jeffrey Breen writes Like Monty Python's Killer Rabbit, cheap indoor antennas seem harmless to satellite and cable providers. But with the digital TV transition in the US, rabbit ears can suddenly provide digital-perfect pictures, many more channels, and even on-screen program guides. Already feeling pressure as suddenly budget-conscious consumers shed premium channels, providers must now get creative to protect their low-end as well.
> Date: Saturday, February 14, @04:55PM
> Tags: business competition usa entertainment tv story

Excerpt 1: A Sample Post.

> Not rabbit ears (Score:5, Informative)
> by Show+Me+Altoids (1183399) on Saturday February 14, @04:57PM (#26858873)
> Rabbit ears don't pick up UHF signals; they are for VHF which is going away. It's the "loop" part of current antennas which will receive UHF.
> * 78 hidden comments

Excerpt 2: A Sample Comment.

Participants in Slashdot discussions may be anonymous or registered users. Anonymous posters suffer a built in penalty of having all their comments start with a score of zero, whereas all registered users start with a score of one. Registered users also have access to a host of additional features on the site such as the ability to become "friends" with other users, set up a personal profile, and obtain the privilege of rating comments to help shape discussions.

Approach

Our approach uses statistical machine learning to gain insight into the mechanisms by which online communities filter and censor content from the bottom up. We began by mining the Slashdot community for features that would allow us to replicate the community rating system. We used a combination of information gain, intuition, and trial and error to identify feature sets that would yield high accuracy.

[2] http://slashdot.org/faq/com-mod.shtml#cm520

[3] http://news.slashdot.org/article.pl?sid=09/02/14/2025245

[4] http://news.slashdot.org/comments.pl?sid=1128309&cid=26858873

We evaluated the features using several machine learning algorithms including neural networks, support vector machines (SVMs), and bayesian approaches. The best results were found using SVMs and the results in this paper reflect this. Ultimately we were able to study the salience of reputation-based, social, and linguistic features to gain insight into the behavior of community filtering as practiced by the Slashdot moderating community.

Features

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[5], a text analysis database designed by psychologists to study the various emotional, cognitive, and structural components in text (Pennebaker, Francis, and Booth 2007). The contextual and author reputation set do not represent the content of the comments. Instead they are based upon contextual information of the comment such as when it was posted or how much discussion it generated, or information about the reputation of the author such as what his or her recent comment ratings. The full list of linguistic, contextual and author reputation features and their descriptions is available on the web[6].

Contextual and Author Reputation Features. These features were primarily based on observations made by sifting through the comment database. For example, we observed that comments that were made a very long time after the original post were much less likely to receive a high rating. This is likely due to the fact that as the day moves on less people are reading the discussion section of old posts. Further features came about in an attempt to exploit the author reputation metrics that exist in the system, such as the average score of the 24 most recent comments by an individual author. Some of these features are specific to the Slashdot community and cannot be directly applied to the rating system for other online communities, but analogous features can be found in many of them. Author reputation features for YouTube, for example, might include how many subscribers they have, how many videos they have posted, and the average ratings of those videos. Some of the features used include:

- *timeDifference* between original post and comment.
- *subComments*: number of replies under the comment.
- *posterIdNumber*: how long a poster has been on Slashdot.
- *posterAcceptanceRatio*: percentage of articles that the user submitted that were accepted and posted as news.

While all of the features we have looked at are useful at gaining insight into how the filtering process works, some features are only in evidence significantly after the comment has been posted (for example, subComments). These ex-post features are indicators of how the community is rating these comments, but they cannot be used in any mixed-initiative system that combines machine learning and collaborative filtering. It is also worth noting that some of the most salient features, such as timeDifference, may represent flaws in the collaborative filtering scheme. It is difficult to say if good comments tend to be timely or if late comments simply do not get the benefit of moderator eyeballs.

Linguistic Features. The motivation behind including linguistic features was our hypothesis that comments that receive higher ratings generally exhibit higher quality writing. We expanded on this with further linguistic analysis based on ideas such as the hypothesis that comments with overall positive sentiment would be more likely to receive a high score. Some of these hypotheses proved to be true, as is explained in the Discussion section. The linguistic features included thirty features from LIWC, seven unigrams, and an additional six features we derived and extracted from the comment text such as the number of words appearing in both the comment and original post.

The advantage to using linguistic features is the ability to easily port them across a variety of rating systems. As the end goal of this research is to create an agent that can augment a collaborative filtering system, effective linguistic features would be an important part of making such an agent as portable as possible. Some linguistic feature include:

- *Comment Sentiment*: Ratio of positive to negative emotion words.
- *Swear Words*: Percentage of swear words in the comment.
- *First Person Pronouns*: Percentage of words that are first person pronouns (i, my, mine).
- *Post Word Count*: The number of words in the comment that also appear in the original news post.
- *Word Count*: The total number of words in the comment.

Evaluation Methodology

All classification was performed using a Support Vector Machine Classifier that used a Gaussian radial basis function. The continuous features were discretized into four bins before classification. A single experiment consisted of generating the data set and feature space, randomly selecting an equal number of comments from each class, and running the SVM classifier through 10-fold cross validation. 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 keeping the class distribution equal as changing the score range for each class affected the maximum number of comments per class. Accuracy measurements were obtained by running each experiment five times, randomly generating a new data set and feature space for each iteration. Classification was performed using the WEKA toolkit[7] and the LIBSVM library[8].

We evaluated the ability of our feature set to predict the community rating of comments made on Slashdot news sto-

[5] http://www.liwc.net/

[6] http://psal.cs.drexel.edu/files/Slashdot_Features.pdf

[7] http://www.cs.waikato.ac.nz/ml/weka/

[8] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

ries on the dates of Saturday, February 14th[9] and Monday, February 16th 2009[10]. We chose these days based on the assumption that the community may behave differently on weekends and weekdays due to the number of people trapped behind a computer monitor for the 9-5 workday. Additionally, we restricted the comments we analyzed to just the first-tier replies, meaning the comments that were direct replies to the original post and not comments that were replies to other comments. We did this based on the assumption that comments made further along in each thread were less likely to be viewed by the whole community and thus less likely to accurately represent the general opinion of the community on what comments were good or bad.

The classification experiments sought to answer a number of questions. What features seem to represent the ways in which the community determines the quality of a comment? Is it possible to predict the original community rating of a comment based on a selection of both linguistic and author and community specific features? Are linguistic features alone useful in determining the quality of a comment? Are "funny" comments more difficult to automatically classify than those which have been labeled "informative?"

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 achieving the overall goal of 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.

Extracting Top Comments

The first task of our classifier was to extract the best comments without attempting to further classify everything else. 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 is an important result despite being relatively straightforward as a classification task because it 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. We looked at modifying our definition of a top comment to be only those with a score of 4 or 5 but found negligible improvements.

Predicting Bad, Neutral, and Good Comments

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." Because of this, It can be important to make a distinction between multiple levels of comment quality. 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.

Actual Class	Bad	Neutral	Good
Bad	324	35	32
Neutral	23	301	67
Good	33	89	269

Table 2: Bad/Neutral/Good Confusion Matrix.

We were able to classify the comments with an overall accuracy of 76%, significantly higher than random-chance classification of 33%. An example confusion matrix for one of the random tests can be seen in Table 2. This confusion matrix shows that in addition to the relatively high accuracy, the misclassifications are skewed in a way that makes sense. For example, almost three times as many "good" comments were classified as "neutral" as were classified as "bad." Since a comment classified as "bad" in the Slashdot community receives penalties such as being automatically hidden it is much more desirable for a "good" comment to be classified as "neutral" instead of "bad."

"Funny" vs. "Insightful" vs. "Informative"

We believed that some posts will be difficult to automatically classify. Humor, for example, is notoriously difficult for automated systems, though there is some promising work on the subject (Dybala et al. 2009). Funny Slashdot comments can be identified by the community in the same way "Insightful" or "Informative" comments are. We used this metadata to determine if "funny" comments were easier or harder to classify. We confirmed our original beliefs by finding that when only "funny" posts are included in the "good" class (since only good posts have the potential for being described as funny), the overall accuracy drops from 76% among 3 categories to 65%. Furthermore, the precision for the "good" class specifically drops from 70% to 52%.

Linguistic vs. Contextual/Reputation Features

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 liguistic 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. 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

[9] http://slashdot.org/index.pl?issue=20090214

[10] http://slashdot.org/index.pl?issue=20090216

Classes	Score to Class Distribution	Feature Set	Overall Precision
2	[-1,0,1,2] [3,4,5]	Linguistic + Contextual + Reputation	82%
2	[-1,0,1,2] [3,4,5]	Linguistic	63%
3	[-1,0] [1] [2,3,4,5]	Linguistic + Contextual + Reputation	76%
3	[-1,0] [1] [2,3,4,5]	Linguistic	42%

Table 1: Overall Accuracy Chart.

yielded an accuracy of 55%. If we removed "funny" comments, however, that accuracy rose to an average of 63%. 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%.

While these numbers are not as significant as our earlier results, they demonstrate that augmenting collaborative filtering with linguistic features that can be extracted across most community filtering systems is possible.

Salient Features. The most salient features all made use of author reputation information, rather than post content. These were features like the number of news posts submitted by the author, the number of friends the author had, the ratio of posts accepted for publication on the site, the length of time the author had been active on Slashdot, and the aggregate score for other comments posted by the author.

Following these features in salience were features related to the properties of the discussion itself, namely the number of subcomments generated by the comment and the promptness of the comment relative to the article being posted.

The most successful set of linguistic features selected were the pronoun-based features, particularly first-person pronouns which indicated a well-received post. We identified other salient linguistic features but they were considerably less effective than pronoun usage. The next most salient features are the length of the comment (longer being better), the number of words the comment had in common with the post, the number of commas, and the lexical density.

Misclassification. There is still considerable work to be done to identify classes of posts that were difficult to classify. However, good short posts were often misclassified, especially when they were posted by an anonymous author. The data suggests, though more analysis is needed, that good posts missed by the classifier often reflect comments of authors with little or negative reputation on the site. In general, anonymous posts were easier to classify than attributed posts as they were more likely to be rated bad.

Discussion

While contextual and author reputation features did provide both good results and insight into the filtering mechanisms of the Slashdot community, there is something unsatisfying about using these features. One goal of a filtering system should be to elevate the good comments of new, rare, or often less useful commenters. While the bulk of good comments may be recognizable by author reputation alone, the value added of community filtering is in recognizing when that is not the case. However, the use of these features show how the structure of a community site and the structured metadata it provides can make this hard classification problem tractable.

Features based on the text alone would allow more democratic filtering and also allow improved filtering of anonymous writing. However, using linguistic features poses a number of difficulties, based on the traditional hardness of natural language processing. Word sense disambiguation was a challenge, as was context. Despite these difficulties, we were able to identify several linguistic features that were salient, showing that determining if a piece of writing is likely to be viewed as "informative" or "insightful" to a community does not necessarily require understanding.

Learning to predict the rating behavior of an online community has identified features that are correlated with high (and low) ratings. Merely replicating the metric, however, does not separate correlation from causation. Are these features truly what the community is looking for in its rating or are these features just correlated with other hidden features that identify good posts? One way to shed light on this is to test if exposing these features to users helps them craft more interesting and better received posts. We plan to perform such user studies in our future work.

Analyzing how frequently certain features appear within good, neutral and bad comments provides insight into the specific mechanisms that the Slashdot community uses when rating a comment. Some features provided more obvious feedback while others supplied surprising insights. For example, one might expect posts that contain more swear words to be ranked poorly and our data supports this claim as swear words appeared 57% more frequently in bad posts than in good posts. This indicates that the more swear words a person uses in their post, the more likely the Slashdot community will give it a lower rating.

Other features, however, produced unexpected results, such as second person pronouns and first person plural pronouns. The results showed that second person pronouns appeared 26% more frequently in bad posts while first person pronouns appeared 34% more frequently in good posts. This could indicate that the Slashdot community rates comments higher if the author of the comment takes ownership of their writing by using first person pronouns instead of using second person pronouns.

Conclusion and Future Work

Information overload is a huge problem on the Internet and moderation requires significant amounts of human labor. Automatic metrics that can help sort through online discourse for insightful or informative content would be useful—even in large communities like Slashdot with complex moderation schemes where many comments do not get tagged or rated. 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—considerably better than 33.3% of chance—achieved using linguistic features alone shows that finding interesting discourse automatically is an interesting and likely achievable goal for natural language processing.

This work also 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.

A number of questions have arisen from our results that pave the path for future work. How well do the features identified in the Slashdot dataset do at replicating trust metrics for other communities? If they are trained on Slashdot data versus data from that community, do other features work better? How do the filtering mechanisms (granularity, democracy, etc) relate to which features are best? What about community demographics?

We already know that short and funny posts are more difficult to classify than longer posts that offer information or insight. However, there is more work to be done in understanding why certain posts are misclassified and whether difficult-to-classify posts can be detected automatically.

Understanding the differences in community filtering standards and procedures could help communities cross-pollinate their discourse. If the community rating mechanisms for two communities could be approximated automatically, then these automated mechanisms could bring relevant content to the attention of new communities. We plan to explore how the features we've identified translate to other communities and if other features or algorithms work better.

We anticipate hard limits to the accuracy of filtering mechanisms based on text-based features that do not actually understand natural language. However, the question of whether one can determine if a piece of discourse will be of interest to a community is not precisely the same as understanding it as has been demonstrated by this research.

References

danah boyd, and Ellison, N. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-mediated Communication* 13(1).

Dingledine, R.; Freedman, M.; and Molnar, D. 2000. *Accountability Measures for Peer-to-Peer Systems*. O'Reilly Publishers.

Dybala, P.; Ptaszynski, M.; Rzepka, R.; and Araki, K. 2009. Humoroids—conversational agents that induce positive emotions with humor. In *Autonomous Agents and Multiagent Systems (AAMAS).*

Farzan, R.; DiMicco, J. M.; and Brownholtz, B. 2009. Spreading the honey: A system for maintaining an online community. In *ACM Conference on Supporting Group Work.*

Finin, T.; Joshi, A.; Kolari, P.; Java, A.; Kale, A.; and Krandikar, A. 2007. The information ecology of social media and online communities. *AI Magazine* 28(3).

Harper, F. 2009. *The Impact of Social Design on User Contributions to Online Communities*. Ph.D. Dissertation, The University of Minnesota.

Herlocker, J. L.; Konstan, J. A.; Terveen, L.; and Riedl, J. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22(1):5–53.

Josang, A.; Ismail, R.; and Boyd, C. 2007. A survey of trust and reputation systems for online service provision. *Decision Support Systems* 43(2):618–644.

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

Lampe, C. 2006. *Ratings Use in an Online Discussion System: The Slashdot Case*. Ph.D. Dissertation, University of Michigan.

Levien, R. 2004. *Attack Resistant Trust Metrics*. Ph.D. Dissertation, University of California, Berkeley.

Massa, P., and Aversari, P. 2007. Trust-aware recommender systems. In *RecSys.*

Newitz, A. 2007. Herding the mob. *Wired* 15(3).

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

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2007. Linguistic inquiry and word count - liwc2007. www.liwc.net.

Resnick, P., and Sami, R. 2008. The information cost of manipulation-resistance in recommender systems. In *ACM Conference on Recommender Systems (RecSys).*

Resnick, P., and Sami, R. 2009. Sybilproof transitive trust protocols. In *ACM EC '09*, 345—354.

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

Terveen, L., and Hill, W. 2001. Beyond recommender systems: Helping people help each other. In *HCI In The New Milenium.*

Xiong, L., and Liu, L. 2004. Peertrust: Supporting reputation-based trust in peer-to-peer communities. *IEEE Transactions on Knowledge and Data Engineering (TKDE), Special Issue on Peer-to-Peer Based Data Management.*