Analyzing and Predicting Community Preference of Socially Generated Metadata: A Case Study on Comments in the Digg Community

Elham Khabiri, Chiao-Fang Hsu, and James Caverlee
Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77845 USA
{khabiri,drakihsu,caverlee}@cse.tamu.edu

Abstract

Large-scale socially-generated metadata is one of the key features driving the growth and success of the emerging Social Web. Recently there have been many research efforts to study the quality of this metadata that relies on quality assessments made by human experts external to a Social Web community. We are interested in studying how an online community itself perceives the relative quality of its own user-contributed content, which has important implications for the successful self-regulation and growth of the Social Web. To this end, we study the community preference for user-contributed comments on the social news aggregator Digg. In our analysis, we study several factors impacting community preference. We propose a learning-based approach for predicting the community's preference rating of unseen comments, which can be used to promote high-quality comments and filter out low-quality comments based on the community's expressed preferences.

Introduction

Socially generated metadata comes in many forms – e.g., tags for annotating Web objects on Flickr and Delicious, user-contributed reviews of books and movies on Amazon, user-contributed comments on blogs, and so on. Increasingly, this metadata is being mined and harnessed for enhanced information filtering, retrieval, and summarization of the underlying object to which the metadata is applied. With the continued growth of the Social Web, a natural concern is on the quality of socially-generated metadata and the potentially negative impact of spam and low-quality metadata on these and other applications for enhanced information access. Indeed, a number of studies have examined the quality of socially-generated metadata across domains; these quality assessments typically rely on experts external to the Social Web community (e.g., a panel of human experts declares that a blog comment is "spam" or "not-spam").

In this paper, we are interested in studying how a Social Web community itself perceives the quality of socially-generated metadata within the community, so that the community is the final arbiter of quality. By studying how a community can self-regulate, we may gain insights into what a community values and how to sustain the positive growth of the community. Concretely, we study the popular social news aggregator Digg and the socially-generated comments that Digg users can annotate news articles with.

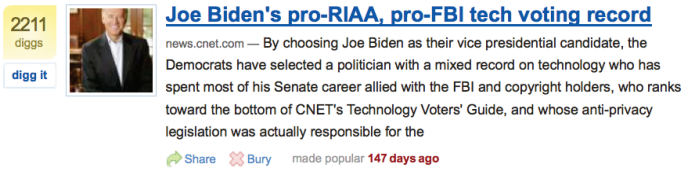

Figure 1: Example story on Digg.

Background and Related Work

Digg is a prominent Web 2.0 news aggregation service in which users can submit stories to the community, rate stories that have been submitted by others (to "Digg" a story is to cast a positive vote for it) and comment on stories. With more than 27 millions visitors in the past year (according to statistics from Compete.com), Digg is one of the most successful social news aggregators among its rivals such as Reddit, Newspond, mixx5, Buzz!Yahoo, and SlashDot.

Figure 1 illustrates an example submission to the Digg community. Our interest in this paper is to study the socially-generated metadata (comments) within the Digg community that has been attached to this news article. Each comment may be rated by members of the community using a simple thumbs-up or thumbs-down rating system. The system aggregates all ratings applied to a comment so that users can filter comments by rating. Comments on Digg range in style and perceived quality within the community; some examples include the informative and highly-rated (like the comment in Figure 2), to the poorly received (see Figure 3).

From a research perspective, Lerman has studied Digg and its article rating system in some detail, e.g., (Lerman 2007). She has shown that users tend to like stories that were submitted by their friends and also were read and liked by them. This reveals that the social network behind Digg plays a significant role of promoting stories to Digg's front page, potentially leading to a tyranny of the minority situation in which a small number of interconnected power users have

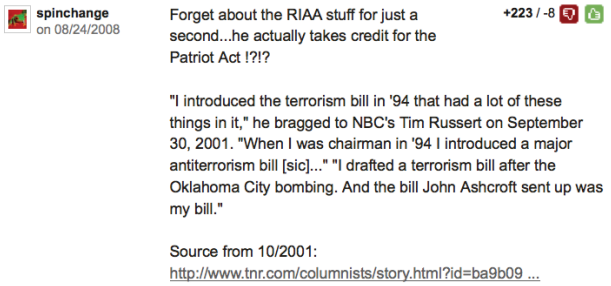

Figure 2: Example highly-rated comment.

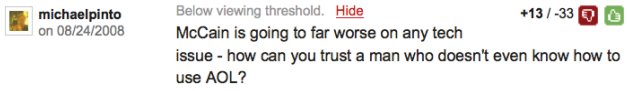

Figure 3: Example lowly-rated comment on Digg.

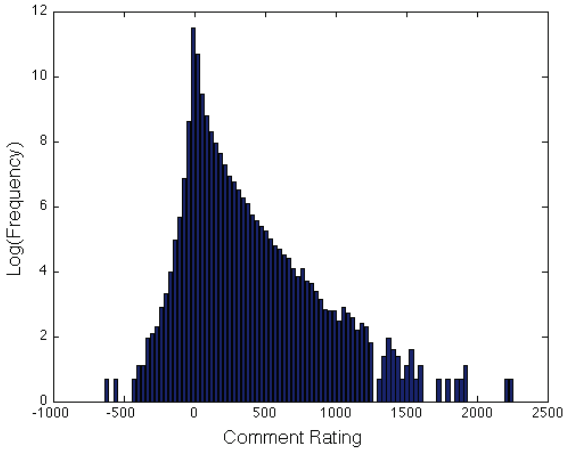

Figure 4: Distribution of Digg comment scores.

the most visibility and influence on the front page. However, to the best of our knowledge, there has been no previous work studying these users and their influence on comments on Digg, nor has there been any general study of Digg comments.

Our work in this paper is inspired by some previous studies of comments in message forums and newsgroups, including (Goldberg & others 1992) and (Mishne & Glance 2006). In particular, the Slashdot community – one of the acknowledged forebears of Digg and related social news aggregators – has attracted much attention. It is important to note that Digg differs from Slashdot in a number of critical dimensions. First, Slashdot offers a restricted form of comment rating (moderation) in which only a fraction of all users are selected to moderate a given comment. This restriction is in direct opposition to the Digg philosophy, in which all users are eligible to rate a comment. Second, Slashdot's comment rating policy restricts the ratings of a comment from -1 to 5, unlike Digg's comment rating system which is (potentially) unbounded, allowing for a wide variety of scores to be applied to comments. The structure of the Digg community could be potentially more problematic for sustaining the growth and quality of the community comment rating system – can the community really rely on the more democratic voting system in which all users can participate? In the rest of the paper, we address this and other research questions in our effort to understand community preference of socially generated metadata.

Data

In November 2008, we crawled the most-Dugg stories of the past 365 days, resulting in a corpus of 4,500 Digg stories containing 232,000 comments submitted by 38,000 unique contributors. We focused our collection on these older pages since the commenting and rating activity has most likely stabilized for these stories, leading to a more reliable analysis of the comments.

We show in Figure 4 the distribution of community ratings for all of the comments harvested from Digg. Note that the majority of comments receive an aggregate positive score, though with some outliers at both the extreme negative and positive ends. The maximum comment score is 2357, the minimum is -861, the mean comment score is 24.27 and the median is 3.

Community Preference for Comments in Digg

In this paper our interest is to study the factors that may influence the community's preference (aggregate rating) for comments on Digg. We begin this study by considering three types of factors that we hypothesize could be important – the visibility of the comment in the community, the influence of the user contributing the comment, and the content itself of the comment.

Visibility of Comment

By visibility of a comment, we are interested to understand the relationship between the size of the community that views the comment and the possible range of scores for the comment. We measure the visibility of a comment through two factors: (i) the Digg score of the article the comment is attached to; and (ii) the order in which the comment has been posted. More visibility of an article implies a higher Digg score for the comments. Figure 5 shows that the mean score of comments that are placed at the beginning of each comment page is greater than the mean score of those comments appearing at the end of the page. This shows that comments that are submitted earlier tend to receive a higher Digg score.

User Reputation and Influence

The second factor we study is the influence of the user contributing the comment. Are there power users in Digg who can attract high comment ratings based on their prestige and position within the social network? To approximate user reputation and influence, we have considered several per-user features:

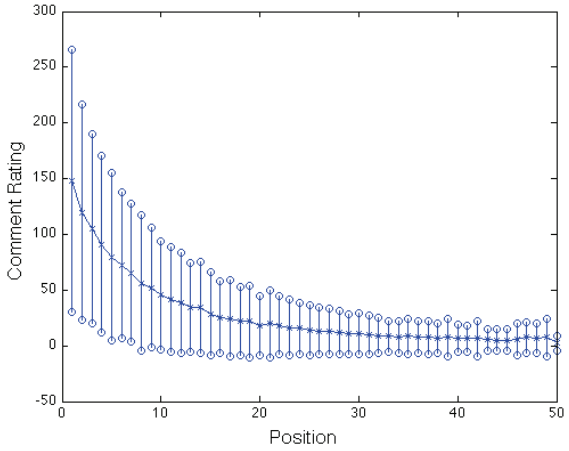

Figure 5: Position of comment versus comment score. We report the mean comment score +/- one standard deviation.

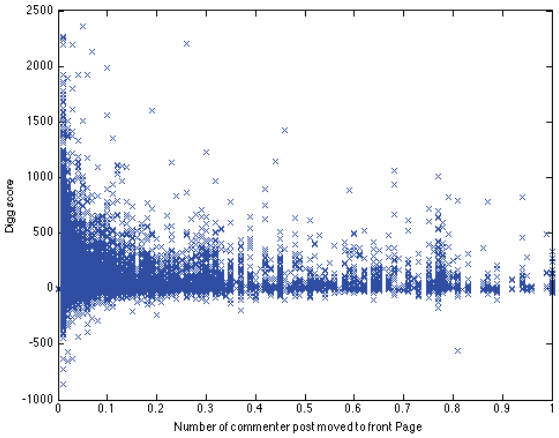

Figure 6: Normalized number of articles appearing on the Digg front page versus comment score.

- *Number of Articles Submitted:* This measures a user's activity in the community by the number of article the user has submitted to the Digg community.
- *Number of Diggs*: This feature is the number of Diggs the user has made on other articles.
- *Number of Articles Appearing on the Digg Front Page*: Digg uses a proprietary promotion algorithm to determine which stories submitted by its users reach the front page of Digg.
- *Number of Profile Views*: How many times has the commenter's Digg profile been viewed? Is this a popular person on Digg?
- *History of Received Comment Ratings*: This feature measures the aggregate (sum) rating of a user's past comments.
- *History of Received Comment Replies:* This feature measures the number of replies that the commenter has received to past comments.

Figure 6 shows the relationship between one of the user-based feature and the comment score. Note that when the number of front page posts by a commenter increases, no increase can be observed for the Digg score for the comment. Similar relationships hold for the other user-based features. Based on these observations in our Digg dataset, it would seem that being an active and influential member of the Digg community is not a good predictor of comment score, though this is open to further study.

Content Analysis of the Comments

The third factor we study are features related to the text analysis of the comment itself. We consider several semantic and statistical features of the comment text:

- *Comment length*: The first feature measures the number of words in the comment text. We hypothesize that the Digg community values average-length comments rather than extremely short or extremely long comments. Although a long comment may be more informative, the community may not appreciate the effort to read and understand it. Studying the relationship between comment score and its length, we found that the comment score is maximum for short comments.
- *Verb/Noun count*: This feature is a simple count of verbs and nouns.
- *Entropy*: The entropy of a comment reflects the richness of the comment by measuring the variety of words in the text. In our experiments we found that comments with less complexity get higher Digg scores. Eq.1 shows that for a text with λ number of words what is the entropy of a text when each of the words has the frequency of p_i.

$$Entropy(text) = \frac{1}{\lambda}\sum_{i=1}^{n} p_i[log_{10}(\lambda) - log_{10}(p_i)] \quad (1)$$

- *Readability:* We measure the readability of a comment by its SMOG score (McLaughlin 1969), which estimates the years of education needed to understand a piece of writing. SMOG considers the number of poly Syllables and the number of sentences in a text. Based on what we observed, comments with higher readability SMOG scores receive higher ratings.

$$SMOG = \sqrt{polySyllables * 30.0/sentences} \quad (2)$$

- *Subjectivity vs. Objectivity*: Subjective comments refer to unjustified personal opinions, in contrast to knowledge and justified belief. We measure the subjectivity/objectivity of each comment using the open source NLP tool LingPipe (Carpenter 2004).
- *Polarity*: Finally, we measure the polarity of each comment using LingPipe (Carpenter 2004) and compare it with the polarity of the article. Our hypothesis is that the community will tend to favor those comments where their polarities matches the polarity of the story.

Predicting Community Preference

Based on our analysis of Digg community preference for comments, we propose a learning-based approach for predicting the community's preference rating of unseen comments.

Prediction Framework

The prediction framework relies on a classification approach for building a predictive model. The goal is to predict for an unseen comment one of four different labels: Excellent, Good, Fair, and Bad. Recall Figure 4, where we plot the distribution of comment ratings in our Digg comment dataset. In our experiment, we define the class boundaries such that a comment with the score of less that -100 is considered as a Bad comment. Comment score between -100 and 0 is Fair, between 1 and 600 is Good and greater than 600 is Excellent. We train two different classifiers over 90% of the comments to build the model using the features described in the previous section. We evaluate the quality of the model over the held-out 10% of the comments. Concretely, the two classifiers used in this paper are Linear regression and Quadratic classifier.

Linear regression Classifier: The relationship between the features is modeled by fitting a linear equation to the ground truth which is the Digg score of the comments in here. Each feature will receive a weight based on the influence it shows on the training data.

$$\sum_{i=1}^{15} f_i * w_i = S \quad (3)$$

where f_i is feature i, w_i is the weight of feature i and S is the Digg score vector of the comments. Through the training process we first obtain the regression weights. Later we apply the learned weight to predict the score for the test samples.

Quadratic Classifier: In a quadratic classifier the posterior probability of each class is evaluated and the class with the largest $P(w_i|x)$ is selected. That is, knowing x as a comment, what is the probability of its membership in class w_i. The class with the highest posterior probability will be assigned to the test sample. With the assumption of Gaussian distribution of the samples the following quadratic equation is used:

$$P(w_i|x) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}log(|\Sigma_i|) + log(P(w_i)) \quad (4)$$

Here μ_i and Σ_i are the mean and covariance of each training class w_i. The prior probability $P(w_i)$ is selected based on the percentages of training set comments in each of these categories.

Preliminary Results

In our initial evaluation, we measure the classification rate, precision, and recall over the test set of comments. The classification rate measures the percentage of the comments that were classified correctly. The precision and recall was calculated for each group (Bad, Fair, Good, Excellent) separately.

Table 1: Results

Method	Rate	Precision	Recall
Reg.	80%	0.01 0.38 0.64 0.00	0.03 0.02 0.94 0.14
Quad.	85%	0.25 0.82 0.70 0.02	0.03 0.23 0.96 0.04

In the Excellent group for example the Precision is the number of actual excellent comments (true positive) retrieved by a our system divided by the total number of retrieved Excellent comments by our system: $Precision = \frac{TP}{TP+FP}$. Table 1 reports the evaluation measures for the two classifiers using the base set of boundary values. We find that the quadratic classifier approach has a higher classification rate as well as higher precision and recall in most groups. We see that the precision for the Fair and Good categories is high (0.82 and 0.70) relative to the precision for the Bad and Excellent categories (0.25 and 0.02). These latter two categories are relatively small and difficult to predict. We are encouraged, however, by the success in differentiating between Fair comments and Good ones.

Conclusions and Future Work

In this paper, we have studied the Digg community and its community preference for user-contributed comments. We have examined the relationship between the comment score and several factors, including the visibility of the comment in the community, the influence of the user contributing the comment, and the content itself of the comment. We have seen that Digg users prefer short, simple, and readable comments, and that so-called power users in the community do not, in fact, wield considerable influence over the scores of comments in the community. Based on these observations, we have proposed a learning-based approach for predicting the community's preference rating of unseen comments. Our initial results are encouraging, and we are extending this work to consider additional features that may improve our success in classifying the extremely high-scoring and low-scoring comments.

References

Carpenter, B. 2004. Phrasal queries with lingpipe and lucene. In *Proceedings of the 13th Meeting of the Text Retrieval Conference (TREC)*.

Goldberg, D., et al. 1992. Using collaborative filtering to weave an information tapestry. *Communications of the ACM* 35(12).

Lerman, K. 2007. Social information processing in news aggregation. *IEEE Internet Computing: special issue on Social Search* 11(6):16–28.

McLaughlin, G. H. 1969. Smog grading: A new readability formula. *Journal of reading*.

Mishne, G., and Glance, N. 2006. Leave a reply: An analysis of weblog comments. In *In Third annual workshop on the Weblogging ecosystem*.