Reporting Incentives in Online Feedback Forums: The Influence of Effort

Radu Jurca and Boi Faltings
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Artificial Intelligence Lab
CH-1015 Lausanne

December 12, 2007

1 Introduction

The internet has made it possible for online feedback forums (or reputation mechanisms) to become an important channel for *Word-of-mouth* regarding products, services or other types of commercial interactions. Numerous empirical studies (Houser and Wooders, 2006; Melnik and Alm, 2002; Kalyanam and McIntyre, 2001; Dellarocas et al., 2006) show that buyers seriously consider online feedback when making purchasing decisions, and are willing to pay *reputation premiums* for products or services that have a good reputation.

Recent analysis, however, raises important questions regarding the ability of existing forums to reflect the real quality of a product. In the absence of clear incentives, users with a moderate outlook will not bother to voice their opinions, which leads to an unrepresentative sample of reviews. For example, Hu et al. (2006); Admati and Pfleiderer (2000) show that Amazon[1] ratings of books or CDs follow with great probability bi-modal, U-shaped distributions where most of the ratings are either very good, or very bad. Controlled experiments, on the other hand, reveal opinions on the same items that are normally distributed. Under these circumstances, using the arithmetic mean to predict quality (as most forums actually do) gives the typical user an estimator with high variance that is often false.

Talwar et al. (2007) analyze other factors that contribute to the user's decision of *when* and *what* feedback to submit to an online forum. They look at hotel reviews from TripAdvisor.com and correlate the numerical ratings with the textual comments left by the reviewers. Their results show that:

- users who comment more on the same aspect of the quality are more likely to agree on a common numerical rating for that particular feature;

[1] http://www.amazon.com

- there is a correlation between the effort spent in writing a review and the transactional risk perceived by the user;
- reviews are biased by the previous reviews submitted by other users;
- users are motivated to submit feedback when they can contribute with new information to the forum;

Improving the way we aggregate the information available from online reviews requires a deep understanding of the underlying factors that bias the rating behavior of users. Hu et al. (2006) propose the "Brag-and-Moan Model" where users rate only if their utility of the product (drawn from a normal distribution) falls outside a median interval. The authors conclude that the model explains the empirical distribution of reports, and offers insights into smarter ways of estimating the true quality of the product.

In this paper we extend this line of research, and investigate the influence of the reporting cost on the distribution of online reviews. Our hypothesis is that higher reporting cost will skew the distribution of reports towards the extremes of the rating scale, confirming thus the intuition behind the Brag-and-Moan model that users with moderate opinions have less incentives to report.

The particularity of this work, however, is to study the relationship between cost and reporting incentives without resorting to controlled experiments, as Hu et al. (2006) do. This objective is challenging at first sight; "public" feedback datasets only reveal the actions of users that did report feedback, therefore nothing can be concluded about the reporting incentives without knowledge of the *total number* of people that should have reported.

Our idea is to use two different sets of reviews about the same items. The first set contains reviews that were elicited through a system A where reporting is very easy. The second set contains the reviews on the same items obtained through a system B, where reporting is significantly more difficult. Since the main difference between systems A and B is the amount of effort required to submit a review, the difference between the distribution of reviews in the two datasets can account for the influence of effort on the reporting incentives.

Several problems might occur when conducting such an experiment. First, we must be reasonably confident that systems A and B address the same community of users. While such a guarantee is normally impossible in any unsupervised online system, we must at least check that the communities addressed by the two systems have similar characteristics. Second, the quantitative effects observed on reporting incentives will only be valid for the difference in effort required by system B as compared to system A. Any wider model that correlates reporting incentives with elicitation effort will have to rely on assumptions about the absolute effort levels required by systems A and B, and on some interpolation scheme that can define the reporting effort of other elicitation mechanisms. Third, since users are not available for post-experiment interviews, we cannot know what other factors were decisive for their observed behavior.

Despite the inconveniences mentioned above, the experimental procedure reported here has a huge advantage: conclusions can be drawn based on the

behavior of tens of thousands of users, something that would hardly be feasible in a controlled experiment setting. In the most pessimistic case, our results can be seen as a very cheap method for guiding more targeted controlled experiments which can fully validate the insights on reporting incentives obtained here. We also hope that this paper can encourage a number of similar initiatives that can *automatically* quantify the behavior of online users by correlating a number of imperfect public databases.

2 Methodology

We look at movie reviews obtained from three different sources:

- IMDB (crawled online from www.imdb.com)
- Netflix (released by Netflix as part of the dataset for the "Million Dollar Prize")
- MovieLens (kindly made available through the GroupLens Research Project from the University of Minnesota)

The IMDB database contains two datasets: (a) the set A of numerical ratings from users who only gave a numerical rating, (b) the set B of ratings from users who also left a textual review for the movie. The two datasets correspond to the two elicitation mechanisms mentioned in the introduction. Leaving only a numerical rating is very easy, and requires only one click from the user (in addition to the effort for visiting the site and finding the movie). Writing a textual comment, on the other hand, is a much more elaborate process. First, users must register with IMDB. Second, IMDB requires at least 10 lines of comments to publish a review. Third, the user must read all the disclaimers regarding spoilers and spoiler warnings.

Hypothesis: Sets A and B will exhibit different probability distributions for the reviews. Since sets A and B are obviously obtained from the same community of users (the ones using IMDB) the difference in the rating distribution can only come from the different effort levels required to submit the reviews.

The next step is to compare the ratings from the IMDB set A to the ratings from Netflix and MovieLens. Intuitively, the ratings from Netflix and MovieLens are the closest we can get to a "complete" rating set. Both systems reward users for rating by giving them improved recommendation. Coupled with the fact that both systems make rating very easy, one would expect almost 0 reporting cost.

Hypothesis: The difference between the rating distributions observed on Netflix and MovieLens, and the ratings from the IMDB set A can account for the effort of visiting a review site and for searching a desired item.

However, when comparing the two datasets we must pay attention to the differences in the communities of users addressed by the two systems. If ratings on Netflix and MovieLens are similarly distributed, one can argue that the community interested in movies and movie ratings has similar characteristics. If, however, the ratings in the two datasets have different distributions, we will

have to see what factors distinguish the two communities, and try to classify the IMDB community somewhere along the same lines.

Finally, the results obtained by comparing the three sets can be used to propose a model for the influence of effort on reporting incentives.

References

A. Admati and P. Pfleiderer. Noisytalk.com: Broadcasting opinions in a noisy environment. Working Paper 1670R, Stanford University, 2000.

C. Dellarocas, N. Awad, and X. Zhang. Exploring the Value of Online Product Ratings in Revenue Forecasting: The Case of Motion Pictures. Working paper, 2006.

D.E. Houser and J. Wooders. Reputation in Auctions: Theory and Evidence from eBay. *Journal of Economics and Management Strategy*, 15:353–369, 2006.

N. Hu, P. Pavlou, and J. Zhang. Can Online Reviews Reveal a Product's True Quality? In *Proceedings of ACM Conference on Electronic Commerce (EC 06)*, 2006.

K. Kalyanam and S. McIntyre. Return on reputation in online auction market. Working Paper 02/03-10-WP, Leavey School of Business, Santa Clara University., 2001.

M. Melnik and J. Alm. Does a seller's reputation matter? evidence from ebay auctions. *Journal of Industrial Economics*, 50(3):337–350, 2002.

Arjun Talwar, Radu Jurca, and Boi Faltings. Understanding User Behavior in Online Feedback Reporting. In *Proceedings of the ACM Conference on Electronic Commerce (EC'07)*, pages 134–142, San Diego, USA, June 11–15 2007.