A Model of Collaboration-based Reputation for the Social Web

Kevin McNally, Michael P. O'Mahony and Barry Smyth
CLARITY Centre for Sensor Web Technologies
School of Computer Science & Informations
University College Dublin
{firstname.lastname}@ucd.ie

Abstract

In this paper we describe a generic approach to modeling user reputation in online social platforms based on an underlying model of collaboration. This distinguishes our approach from more conventional reputation models which are often based around ad-hoc activity metrics. We evaluate our model with respect to a conventional reputation model used by 3 social Q&A websites, each based on a different topical domain.

Introduction

The social web reflects an important paradigm shift in the nature of our online transactions. We increasingly rely on the views and opinions of others to mediate these transactions and as such the reliability of these users becomes an important indicator of quality. Thus, concepts like trust and reputation have become increasingly important in the context of today's social web. It is not surprising that there has been considerable recent research on various ways to measure, quantify, and evaluate the trustworthiness and reputation of users as they participate in a diverse array of online interactions (Resnick and Zeckhauser 2002; Cheng and Vassileva 2005; Recuero, Araujo, and Zago 2011).

In this paper, when we refer to *trust* we are referring to an asymmetric relationship between two users: user *A* can *trust* in user *B* to a lesser or greater extent. Trust refers to the feedback provided by one user in the context of some interaction with another user. For instance, on eBay a buyer can rate a seller and this corresponds to a degree of trust that the buyer has in the seller. When we refer to *reputation* we are referring to a measure that is associated with an individual user, generally as some function of the individual trust scores that have been assigned to this user. For example, an eBay seller might have an average rating of 90%, which corresponds to their reputation in the sense that it is a measure of their trustworthiness across multiple transactions. The idea of distinguishing online trust from reputation has been explored by (Mui, Mohtashemi, and Halberstadt 2002).

In previous work we have proposed a collaboration-based model of reputation for the social search utility HeyStaks (McNally et al. 2011). The key contribution of this paper is the examination of the idea that such a model generalizes to other domains, in this case, Social Q&A. Recent work indicates that a big motivation behind applying reputation systems is to incentivize users of a platform to engage online with others. This can be aided by users trusting the system (Joinson 2008), but also by rewarding users with positive feedback (Cheng and Vassileva 2005) or with improved social standing (Recuero, Araujo, and Zago 2011). A successful reputation system can not only distinguish trustworthy users from untrustworthy ones, but also determine the quality of the resources users provide. Many of these systems calculate trust or reputation by building a network based on some explicitly or implicitly gained information in a specific context. To date, work has focused largely on developing trust/reputation models in a specific setting or context. For example, (Hong, Yang, and Davison 2009) examined modeling reputation on social Q&A sites by leveraging the links that exist between users based on "best answer" feedback. In this paper we intend on extending a generic model of collaboration-based reputation to the same domain. We take a principled approach to calculating reputation. In particular, we consider the following criteria in our model:

1. *What is the structure of the model?* In every case we can model an online community as a graph that makes explicit the interactions between *producers* and *consumers* of information. For example, in a social Q&A setting, which we will examine in detail in this paper, interactions between users can be viewed as users voting on each other's produced answers.

2. *What are the fundamental units of collaboration between users?* The answer to this question is dependent on the domain being examined. In social Q&A, collaboration can be viewed as one user voting on another user's answer to a question posed by a member of the community, resulting in an edge being drawn in the graph between the pair of users. We refer to such a singular instance of interaction as a *collaboration event*.

3. *How do we aggregate units of collaboration to calculate reputation?* How can we leverage the level of collaboration occurring between members of an online community to deduce their reputation? This can be achieved using any number of node aggregation techniques.

A Computational Model of User Reputation

Our model of reputation is based on naturally occurring *collaboration events* that are frequent in many different social web settings. These events can be associated with trust scores to reflect the quality of interaction between collaborating users. Collections of such events naturally form a collaboration graph, and trust scores that flow between pairs of users can be aggregated to evaluate the reputation of individual users. It has been shown in previous work that user reputation information can be utilized to positively influence recommendations made by the social search utility HeyStaks (McNally, O'Mahony, and Smyth 2011; McNally et al. 2011). Here we show how this model can be generalized to other domains, focusing on calculating user reputation as a means of discovering the quality of users' answers on the Stack Exchange network. We evaluate this model by measuring its correlation to a ground-truth metric, and comparing our own reputation scoring metric - "WeightedSum" (WS) to that of standard PageRank (PR) (Brin and Page 1998), as outlined below. For more details on these metrics and how they are employed in a reputation scenario, see (McNally, O'Mahony, and Smyth 2011). Finally, we compare both these approaches to calculating reputation with scores calculated by Stack Exchange (SE), using their own activity-based reputation framework.

Reputation in Social Q&A Sites

The Stack Exchange Network[1] is a popular group of social Q&A sites. There are currently 83 different topical Stack Exchange websites, hosting almost 2 million users. Some 3.8 million questions have been posed, eliciting 7.7 million answers. Users are permitted to post questions that can be answered by other users in the community. Each answer given can be voted up or down by others and the questioner can chose to highlight a single answer as correct; indicating the question has been answered satisfactorily or that answer was the best answer provided.

Collaboration-based Reputation on Stack Exchange

The reputation model proposed can be applied to Stack Exchange data, assuming that a suitable collaboration event can be defined. In this particular scenario we have chosen to model collaboration events as instances of question-answers with individual question-answer pairs scored according to the relative proportion of votes received.

On Stack Exchange sites, users can give feedback on questions and answers by positively or negatively voting on them, and these votes are aggregated by summation. For the purpose of our collaboration events we only include answers that have received a positive aggregate score; an edge is drawn between two users only if the user's answer has received more positive than negative votes from their community. In this case of a given set of question-answer pairs it leads to the partial collaboration graph shown in Figure 1. In turn each collaboration event is scored based on the proportion of votes that the corresponding answer has received.

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

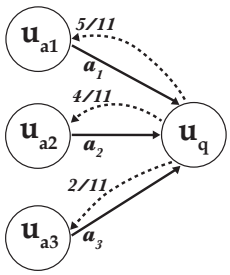


Figure 1: The collaboration event and trust scores corresponding to the example set out opposite.

Using this voting data, we define the following collaboration event. For each answer that received an aggregate vote greater than 0, the associated producer (i.e. the user who provided the answer) receives a trust score in proportion to the vote received. For example, suppose a question is posed on the site and receives five answers, each of which receive an aggregate vote score of +5, +4, +2, -1, -2, respectively. Let $u_{a1}, u_{a2}, \ldots, u_{a5}$ denote the producers of these five answers and let u_q denote the consumer who posted the question. After eliminating answers with negative net votes, the unit of trust scores are divided between producers u_{a1}, u_{a2} and u_{a3} in the proportions $\frac{5}{5+4+2}$, $\frac{4}{5+4+2}$ and $\frac{2}{5+4+2}$, respectively; see the edge weights in Figure 1.

Once the collaboration graph has been created, by adding edges for all of the collaboration events in a given Stack Exchange dataset/domain, the reputation of each user can be calculated according to WS and PR. Briefly, WS is a count of all weights on incident edges into a given (producer) node; note a user who is part of the graph but has no incident in-links will not receive a score (see (McNally, O'Mahony, and Smyth 2011) for details). This is unlike PR, where all connected users receive a small default score.

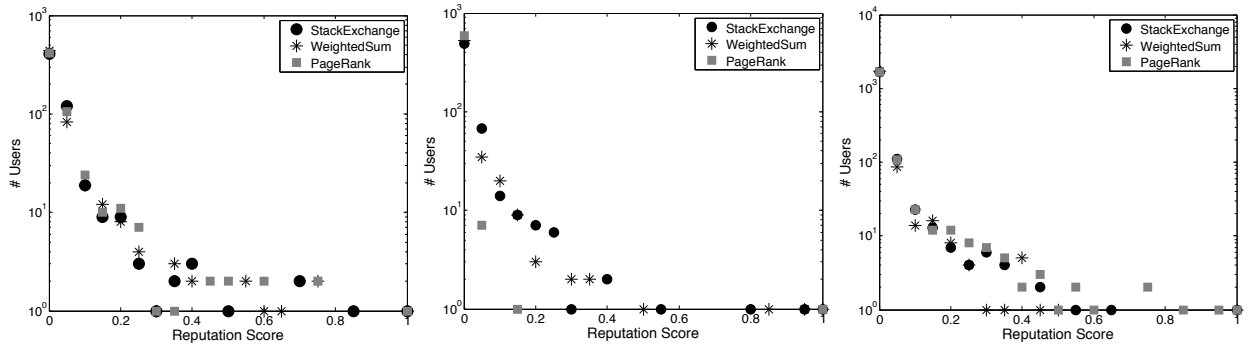
Once the reputation of users has been calculated, it can then be leveraged for a variety of purposes; for example, to re-rank answers to questions based on the answerers' reputation or to provide a means to automatically route questions to users with high reputation in the domain. In this paper, we focus on computing user reputation and leave such applications of reputation to future work.

Evaluation

The current reputation model in Stack Exchange is a typical conventional model based on an aggregation of various user *activities* on the site, and as such it does not represent a 'pure' model of user reputation. Further, it is a domain-specific approach, requiring a fine-grained weighting of the significance of the various activities users can perform on the site. In this evaluation we compare our collaboration-based reputation model, which assumes little domain knowledge, to this conventional approach.

Dataset and Methodology

For this evaluation we selected three Stack Exchange sites, crawling each site to download a complete picture of user activity. These datasets are summarized in Table 1.



(a) Seasoned Advice (b) Electrical Engineering (c) Mathematics

Figure 2: Histograms showing distribution of scores according to the three reputation models for (a) Seasoned Advice, (b) Electrical Engineering and (c) Mathematics datasets.

	Dataset		
	SA	EE	Math
# Users	7,552	7,692	22,242
# Questions asked	5,135	7,783	35,251
# Answers	14,687	18,900	57,146
# Correct Answers	3,456	4,863	24,978
# Votes on Answers	48,947	56,134	220,856
# Users with >0 Questions Answered Correctly	579	604	1,845

Table 1: Summary statistics for three Stack Exchange datasets: Seasoned Advice (SA), Electrical Engineering (EE) and Mathematics (Math).

To conduct our experiment we only considered questions that received at least one answer and where answers have received at least one vote from the community. The smallest of the three datasets examined is from the Seasoned Advice website, dealing with topics on cooking. A similarly sized dataset is taken from the Electrical Engineering website, and the third and largest dataset is from the Mathematics website. We chose the three datasets based primarily on the nature of questions asked on each site. For example, the Cooking dataset provided us with answers to questions that can be more qualitatively assessed by the community. Conversely, answers given on the Mathematics website tend to be either right or wrong, although answers can be voted on with the manner of delivery in mind. The Electrical Engineering dataset provides both answers that can be interpreted as right or wrong, and answers that can be assessed qualitatively.

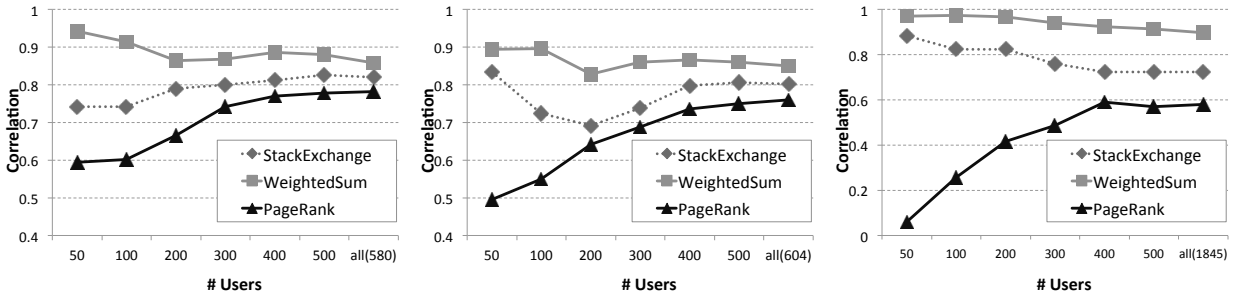
We compare the WS and PR variations of our collaboration-based model by building a collaboration graph for each dataset as previously described. Figure 2 shows histograms which illustrate the distribution of reputation scores across each of these models and datasets. Distributions of user reputation according to the conventional SE model are also shown. For clarity, we have normalized each user's reputation score by the maximum score found in each dataset, according to the individual reputation model presented. All charts indicate the long-tailed nature of user reputation. Each reputation model tends to distribute reputation scores in a similar way for each dataset, although proportionately more users receive a score between 0.05 and 0.1 in the Seasoned Advice dataset compared to the other two datasets.

In order to evaluate these three reputation models we establish a ground-truth as the basis for comparison. In the Stack Exchange network each questioner has an opportunity to mark a single answer as *correct*. This is clearly a strong indicator of answer quality and it is reasonable to consider users who are associated with many correct answers to be more reputable than users who are associated with fewer (or no) correct answers. Table 1 shows the number of users who have answered at least one question correctly according to the questioner. As such, for the purpose of this evaluation, we will use the number of correct answers for a user as a ground-truth for their reputation. And by correlating the reputation scores from the three models with respect to these ground-truth scores we can analyse the performance of each model; under the assumption that higher correlations are to be favoured because they indicate a given reputation model as a stronger predictor of the ground-truth.

Correlation Analysis

Ultimately the true test of our evaluation models is the extent to which reputation correlates with our evaluation ground-truth; in this case the number of correct answers provided by a user. Figures 3(a)–3(c) show the correlations between each reputation model and the ground-truth for different sized groups of users in descending order of reputation. In each chart the maximum number of users corresponds to the number of users for whom ground-truth information is available. For example, in Figure 3(a) we can see that for the top-50 most reputable users in the Seasoned Advice dataset, WS enjoys a correlation coefficient of 0.95 with the ground-truth compared to 0.75 and 0.6 for SE and PR, respectively.

Overall we can see that the correlation achieved by the WS model tends to degrade as we consider larger sets of users with decreasing reputation values. This is to be expected. As we verge towards users with lower levels of activity a small difference between two users' reputation scores becomes more pronounced, having a greater effect on their rank, and thus the correlation. In contrast, the correlation of the PR model tends to increase as we consider larger sets of users with decreasing reputation values. In this regard, the assignment of default reputation scores to users by the PR model has a positive effect on the ranking of users with small



(a) Seasoned Advice (b) Electrical Engineering (c) Mathematics

Figure 3: Correlations between the ground-truth and the three reputation models, for each of the three datasets.

levels of activity, leading to higher correlations in our experiments. In addition, the effect of propagating scores featured in the PR model is likely a primary cause of the degradation in rank correlations between the top 50 or 100 users and the ground-truth. For example, the top ranked user in the Electrical Engineering dataset according to PR is ranked second by the other two models and, indeed according to our ground-truth. Similarly in both the Mathematics and Seasoned Advice datasets, PR is good at identifying the top 50 users in accordance with the ground-truth, however their ranks are considerably different. Another feature of PR that may be having a negative effect is the fact that an individual's reputation influences that of adjacent users. This idea has been explored in (Hong, Yang, and Davison 2009).

In Figures 3(a)–3(c) the WS model consistently outperforms SE, which in turn outperforms PR. We can then conclude that our WS model performs very well with respect to the ground-truth, delivering effective reputation estimates per user and importantly without the need for fine-grained tuning or deep domain knowledge.

Conclusions

The key contribution of this work is the presentation of a principled approach to modeling user reputation in online social platforms. Collaboration between users can be leveraged to measure the reputation of users in potentially any online collaborative environment. We present this approach to calculating reputation as an alternative to more traditional activity-based models. In particular we have compared two variations of our model to the current Stack Exchange reputation model using data drawn from three diverse Stack Exchange Q&A domains. In each case we examined the correlation between user reputation and the number of correct answers provided by users as a ground-truth to demonstrate superior performance for our WS variation when compared to the SE and PR alternatives. This analysis is not without its limitations, particularly given that only one of many possible interaction scenarios was examined. In the future it would be useful to analyze a reputation model against other types of interaction and, indeed, examine the effect on reputation of building the collaboration graph using different kinds of activity.

Acknowledgments

This work is supported by Science Foundation Ireland under grant 07/CE/I1147.

References

Brin, S., and Page, L. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *WWW '98: Proceedings of the 7th International Conference on World Wide Web*. ACM.

Cheng, R., and Vassileva, J. 2005. Adaptive Reward Mechanism for Sustainable Online Learning Community. In *Proceedings of the 2005 conference on Artificial Intelligence in Education*. IOS Press.

Hong, L.; Yang, Z.; and Davison, B. 2009. Incorporating participant reputation in community-driven question answering systems. In *Computational Science and Engineering, 2009 (CSE'09)*, volume 4, 475–480. IEEE.

Joinson, A. N. 2008. Looking at, Looking up or Keeping up with People?: Motives and use of Facebook. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, 1027–1036. ACM.

McNally, K.; O'Mahony, M. P.; Smyth, B.; Coyle, M.; and Briggs, P. 2011. A Case-study of Collaboration and Reputation in Social Web Search. *ACM TIST: Transactions on Intelligent Systems Technology* 3(1).

McNally, K.; O'Mahony, M. P.; and Smyth, B. 2011. Evaluating User Reputation in Collaborative Web Search. In *3rd Workshop on Recommender Systems and the Social Web, Recsys 2011*. ACM.

Mui, L.; Mohtashemi, M.; and Halberstadt, A. 2002. A Computational Model of Trust and Reputation. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 7 - Volume 7*, HICSS '02, 188–. IEEE Computer Society.

Recuero, R.; Araujo, R.; and Zago, G. 2011. How Does Social Capital Affect Retweets? In *The International AAAI Conference on Weblogs and Social Media (ICWSM '11)*. The AAAI Press.

Resnick, P., and Zeckhauser, R. 2002. Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System. *Advances in Applied Microeconomics* 11:127–157.