The Annotation-Validation (AV) Model: Rewarding Contribution Using Retrospective Agreement

Jon Chamberlain
School of Computer Science and Electronic Engineering
University of Essex
Wivenhoe Park, CO4 3SQ England
jchamb@essex.ac.uk

ABSTRACT

Evaluating contributions from users of systems with large datasets is a challenge across many domains, from task assessment in crowdsourcing to document relevance in information retrieval. This paper introduces a model for rewarding and evaluating users using retrospective validation, with only a small gold standard required to initiate the system. A simulation of the model shows that users are rewarded appropriately for high quality responses however analysis of data from an implementation of the model in a text annotation game indicates it may not be sophisticated enough to predict user performance.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User/Machine Systems—*Human information processing*

Keywords

crowdsourcing, gamification, user reward model

1. INTRODUCTION

Evaluating contributions from users of systems with very large datasets is a challenge across many domains, from task assessment in crowdsourcing to document relevance in information retrieval. With a set of tasks with known solutions (called a gold standard) users can be assessed, however these resources tend to be small, if they exist at all. The majority of contributions will not have a known answer at the time the user inputs the data so it is a challenge to appropriately reward the user for a high quality response and distinguish it from unwanted noise.

New users may initially perform badly but should improve with training and experience although lapses in concentration may still cause dips in performance. All users should be encouraged to contribute as the "long tail" of a collaborative data collection effort may account for as much as 30% [6].

GamifIR'14, 13 April 2014, Amsterdam, Netherlands.

http://dx.doi.org/10.1145/2594776.2594779 .

This paper introduces a model for rewarding and evaluating users using retrospective validation, with only a small gold standard required to initiate the system. The model is tested using a simulation and compared to real data from an implementation in a text annotation game.

2. RELATED WORK

The concept of using game elements within a non-game context has a long tradition but only recently has the term "gamification" been defined [5]. Feedback can be given to the user by tracking their performance in the system in order to encourage higher quantity or quality of work and motivational rewards can then be applied. Leaderboards and other comparative techniques show how well users are performing against their peers. User assessment in leaderboards can also be used as competency models, taking a multi-dimensional view of the user's abilities at different tasks [12]. By using such methods gamification aims to change the user's behaviour to meet the goals of the system designers [14].

One example of gamification in information retrieval is asking users to judge the relevance of documents retrieved from a search query using different search algorithms [7]. Whilst this demonstrates motivating a group of users to do this task, it has been criticised for being limited by situational relevance, i.e., that what is relevant to the user that makes the query may not be relevant to the user that is judging the relevance [11]. Using implicit data that is generated by the user in the completion of their task may be a better way to assess relevance [1].

Taken to its extreme, gamification becomes an approach more like games-with-a-purpose (GWAP), where the task is entirely presented as a gaming scenario rather than as a task with gaming elements applied. Many different GWAP have been produced [3] however there has been little research on user assessment frameworks employed within them [13].

When aggregating the data it is no longer sufficient to use simple majority voting as malicious contributions can outweigh real user's responses. Weighted voting or validation are needed to ensure that a contribution is from a "real" user, to assess their understanding of the task and predict whether they will provide a good solution. The superuser reputation scoring model in the social gaming network foursquare[1] hints at the considerable commercial interest in assessing user contributions and similar models are employed by other crowd-based QA datasets such as Stack Overflow [2].

[1]http://engineering.foursquare.com/2014/01/03/the-mathematics-of-gamification

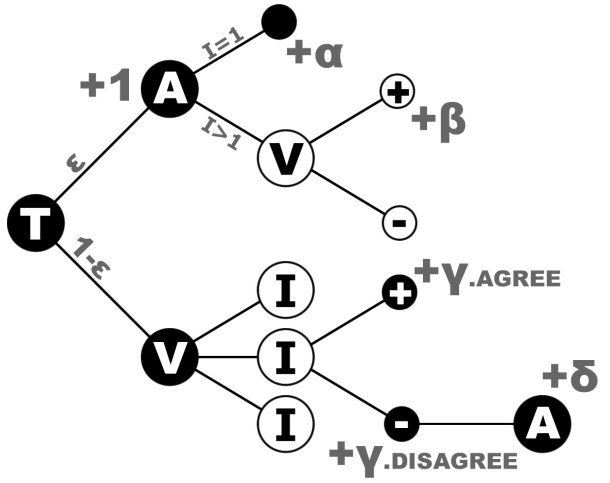

Figure 1: A representation of the AV Model showing how a score is calculated for a user task (T) in either Annotation Mode (A) or Validation Mode (V). Black circles indicate a user input and white circles indicate an input created by others users in the system (the user base).

3. THE AV MODEL

The Annotation-Validation (AV) Model describes a shift from effort-based reward, where the reward is proportional to the number of tasks completed irrespective of the quality, to agreement-based reward where the users receive more reward for higher quality (or more commonly agreed with) solutions. There are 3 scenarios for the model that are considered in this paper:

1. to reward users appropriately for solutions to tasks without assessing quality with a gold standard;
2. to assess user ability by predicting their response to the tasks;
3. to filter a noisy dataset with post-processing.

Initially a user completes an annotation task (Annotation Mode) and is given a fixed reward for their contribution. If the initial group of users (U_A) enter the same solution they are all rewarded again (α), however it is likely they will create multiple solutions or interpretations (I) for the task. In the latter case each interpretation is presented to further users (U_V) in a binary (agree or disagree) choice task (Validation Mode). The validating user is rewarded for every annotating user that they agree with (γ). If they disagree with the interpretation they receive a reward for every annotating user that entered a different interpretation to the one presented, hence they must also be disagreeing. If the validating user disagreed with the interpretation they are asked to enter a solution via Annotation Mode and are rewarded again for their contribution (δ). If the user creates a new interpretation this will also be validated. Every time a validating user agrees with an interpretation, any user from the original annotating group that entered the interpretation also receive a retrospective reward (β).

Additionally, P_u is the probability that the user selects the correct answer (also called the user rating) which is calculated by giving the user a small set of tasks with a known answer; P_{ub} is the mean probability of a user in the system (the user base) selecting the correct answer; ϵ is the proportion of tasks presented in Annotation Mode; and S is the predicted score per task for the user.

$$\alpha = P_u P_{ub}^{U_A-1} + \frac{(1-P_u)(1-P_{ub})^{U_A-1}}{(I-1)^{U_A-2}}$$

$$\beta = U_V(1-\alpha)(P_u P_{ub} + (1-P_u)(1-P_{ub}))$$

$$\gamma = \frac{U_A(P_u P_{ub}+2(1-P_u)(1-P_{ub})+P_u P_{ub}(I-1)+(1-P_{ub})(I-2)}{I}$$

$$\delta = \frac{1-P_u+P_u(I-1)}{I}$$

$$\epsilon = \frac{U_A}{U_A+U_V I}$$

$$S = \epsilon(1+\alpha+\beta)+(1-\epsilon)(\gamma+\delta)$$

The model makes several assumptions (the implications of which are discussed later):

- I is greater than 1;
- there is only 1 correct interpretation per task;
- the user will try to solve the task by choosing the correct interpretation;
- the user only sees the task once.

4. SIMULATION

The AV Model is simulated to predict a score per task (S) for a user of a given rating (P_u). For all the simulations and implementation there were 8 annotating users per task (U_A=8) and 4 validating users per interpretation (U_V=4).

4.1 Task difficulty

The difficulty of the dataset will have an impact on the number of interpretations (I) that are submitted by the users, with more difficult tasks having more interpretations. The score per task in Annotation Mode does not seem to be affected by the difficulty of the dataset, with high rated users only scoring slightly more. This is because each annotation is presented to the same number of validators irrespective of the number of interpretations for the task. The score per task in Validation Mode is different between levels of difficulty, with harder tasks scoring more for higher rated users (see Figure 2).

4.2 Quality of the crowd

A measure of how well the users of the system (the user base) are performing generally is essential when using a validation method. The system increases the score of an annotation using validations so if the users that are validating are not performing well this could have a negative impact, not only on the data quality but also on the motivation of the users. In three different scenarios of user base rating (55% as near chance; 75% as an average response; and 95% as a good response) the model performs correctly, i.e., highly rated users score more per task than poorly rated users (see Figure 3). This effect is magnified when the user base is very good but the model still rewards appropriately even when the user base rating is poor (close to chance).

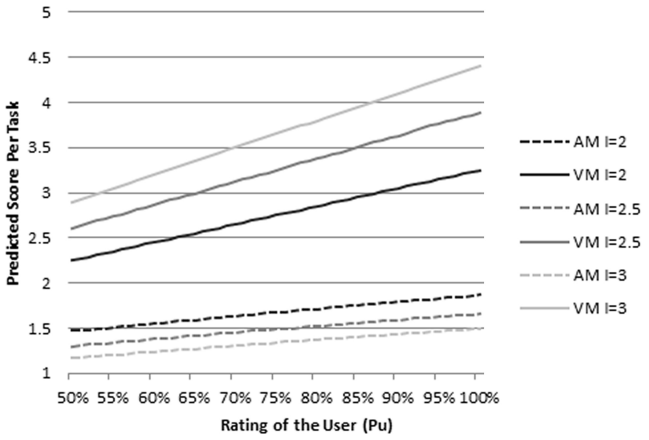

Figure 2: Simulation of score per task for different user ratings, comparing Annotation Mode (AM) and Validation Mode (VM) with different interpretations (I) per task (P_{ub}=0.75).

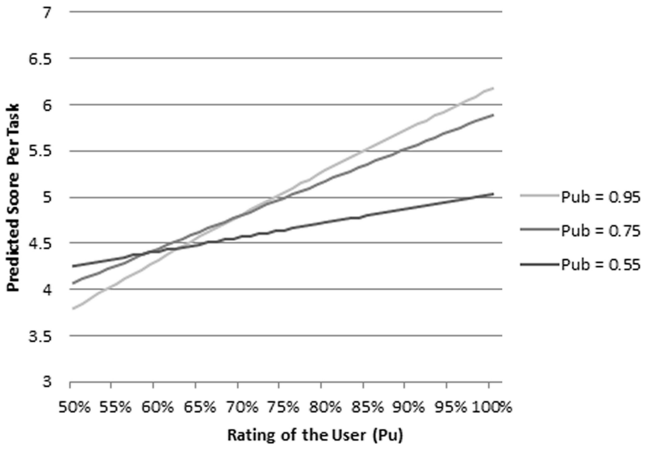

Figure 3: Simulation of score per task for different user ratings, comparing different ratings (P_{ub}) for the user base (I=3).

4.3 Data maturity

During the lifecycle of data being annotated with the model the user will be presented with different proportions of annotation tasks compared to validation tasks (ϵ). When the data is initially released the user will be given annotation tasks (ϵ=1). As more annotations are collected the number of validations presented to the user increases until all tasks have been sufficiently annotated and now only require validations (ϵ=0).

Higher rated users will score more per task and this increases as more validations are required (see Figure 4). This is due to higher rated user's annotations being agreed upon more by validators and thus should increase the motivation of users as the data matures.

The simulation of the AV Model shows that theoretically users can be rewarded appropriately using retrospective agreement for tasks where the solution is not known and users should be motivated to provide higher quality solutions to increase their reward.

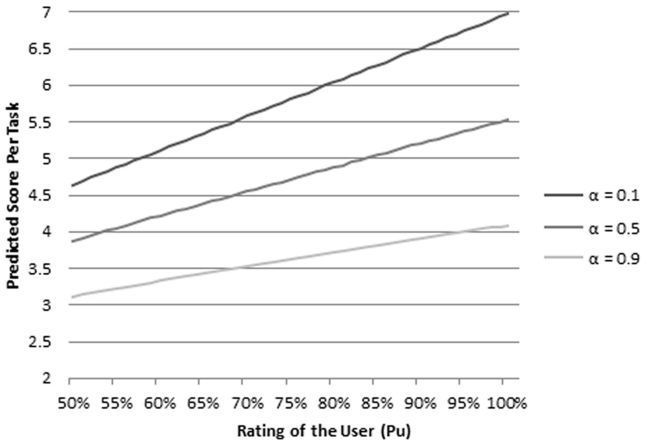

Figure 4: Simulation of score per task for different user ratings at different stages of data maturity (I=3 and P_{ub}=0.75).

5. IMPLEMENTATION

The AV Model was implemented in Phrase Detectives[2] [3], a game-with-a-purpose designed to collect data on anaphoric co-reference[4] in English documents [8].

The game uses the 2 modes to complete a linguistic task: Annotation Mode (called Name the Culprit in the game - see Figure 5) where the player makes an annotation decision about a highlighted section of text (called a markable) and Validation Mode (called Detectives Conference in the game - see Figure 6). Additionally there are gamification features to motivate user contribution, such as leaderboards and high score awards.

The first analysis uses coarse game data (total task-based score divided by the total number of tasks completed per user) and shows a significant correlation between score and user rating (n=1,329, p<0.05, Pearsons), with a similar slope gradient to the model when simulated using P_{ub} and I from the data (see Figure 7). However this data is incomplete as users may not have collected all the points for their work. Also, users should only be able to score a maximum of 9 points per task (full disagreement in Validation Mode and then making a correction in Annotation Mode) however a feature of the game is that a user can skip or cancel the tasks they have been given. Any rewards from cancelled tasks is kept by the user but not included in this calculation which explains the outliers. However, at this coarse level it appears that users are rewarded in a way that the model would predict.

The second analysis only used tasks from completed documents from the Facebook version of the game. Here the user's rating was frequently tested and recorded with each task so it was a more accurate measure of the user's ability as they progressed through the game. Additionally all annotations and validations would have been collected (see

[2]https://anawiki.essex.ac.uk/phrasedetectives
[3]https://apps.facebook.com/phrasedetectives
[4]Anaphoric co-reference is a type of linguistic reference where one expression depends on another referential element. An example would be the relation between the entity *'Jon'* and the pronoun *'his'* in the text '*Jon rode his bike to school.'*

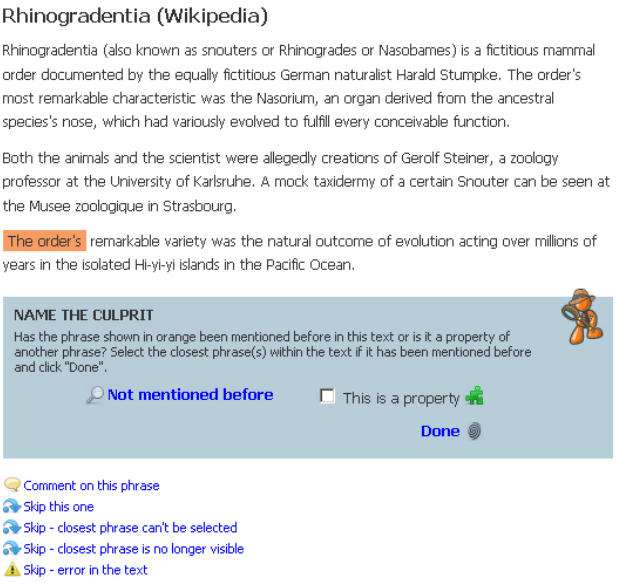

Figure 5: A task presented in Annotation Mode.

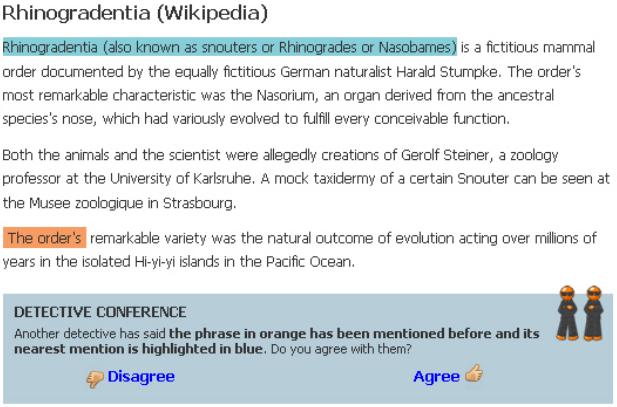

Figure 6: A task presented in Validation Mode.

Figure 8). This data did not show a correlation between score and user rating (n=65,528, Pearsons). This may be an indication that the model is not sophisticated enough to predict a user score and there are confounding factors such as annotation time, gender, system interface, and task difficulty.

6. DISCUSSION

The AV Model makes some assumptions about the data and how the users will interact with it.

Whilst hypothetically possible to have a value of I=1, i.e., only 1 interpretation per task, there would be no value in using a system like this as all the users would enter the same decision, either because the task is very easy or the users are very good.

The model assumes there is only 1 correct solution but in the case of linguistic analysis, relevance judgement and many other applications there are likely to be more than 1 possible

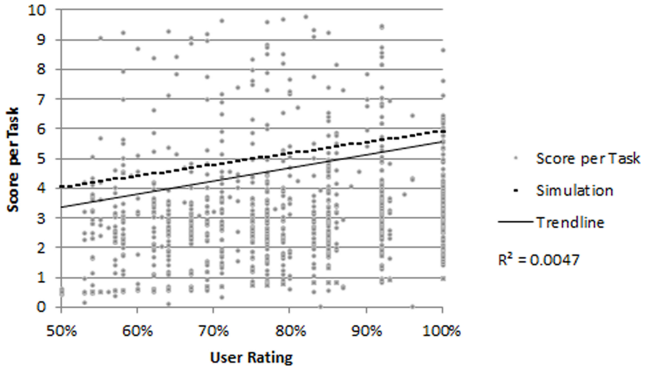

Figure 7: User-based correlation of score per task and rating, not showing outliers with more than 10 points per task (P_{ub}=77.9 and I=2.3).

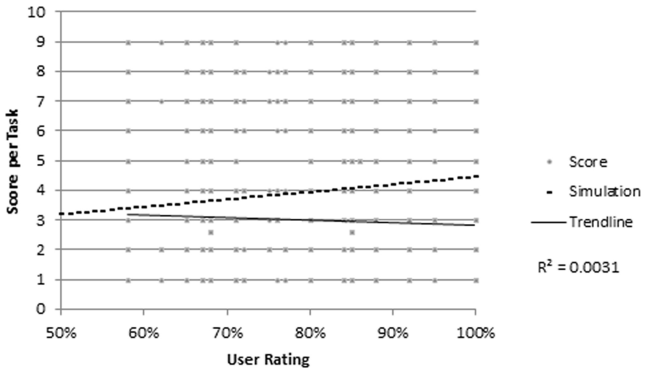

Figure 8: Task-based reconstruction of scoring from Phrase Detectives on Facebook (P_{ub}=78.3 and I=3.1).

answer and the model should be extended to accommodate multiple correct interpretations. Interpretations added after the initial group of annotators have submitted their solutions allows the system to capture less popular solutions and avoid convergence, where users choose what they think will be a popular solution, rather than the best solution.

It is assumed that the user will always try to select the best solution but this is clearly not the case for some users who employ strategies to maximise rewards for minimum effort. There are numerous ways a user can manipulate a system to their advantage and it is the job of system designers to minimise this impact, either at the moment of entering the data or in post-processing.

One strategy identified from the Phrase Detectives game was to enter the fastest and most predictable combination of inputs in order to gain points by quantity rather than quality. Post-processing of this noisy data is required by looking at performance measures such as the time to complete a task [4]. There is also the possibility that users can collude in their answers as it is in their best interest to agree with each other however in Phrase Detectives users could not communicate directly.

The model assumes that the user only receives the task once, in either mode, however in Phrase Detectives this was not the case. Users were occasionally given the same task (although not necessarily in the same mode) in order to com-

plete documents. This may be a source of bias but in practice it happened rarely. When users complete the same task more than once it is possible to measure implicit agreement, i.e., the probability the user provides consistent results. As the user's ability improves over time they may provide different, higher quality answers to tasks they have done before and this could be used to normalise their result set.

Users are rewarded for agreement and not punished for being disagreed with. Scoring models of this kind do exist [9] however it seems intuitive that positive behaviour is reinforced in crowdsourcing. The social network Facebook has resisted repeated calls from users to add a dislike button for presumably this reason, especially as their content is linked to advertising. It may be that negative scoring would produce better results when using the model in post-processing or if the user didn't know they were being punished.

This investigation only used a fixed number of annotation and validation tasks, which were determined by research prior to the release of the Phrase Detectives game. It would be beneficial to reduce the amount of work required per task without reducing the quality to make the most of the limited resource of user contribution.

The AV Model could be used in combination with a recommender system to improve baseline relevance scores in IR and retrospectively weighted scores for incoming and outgoing links may improve graph structures. Using this model in conjunction with sophisticated crowd analysis techniques [10] may yet show that it can predict a user response and measure user performance on a per task basis. Whether user responses can be evaluated in this way at the time of data entry to provide feedback to the user (and a score) presents issues of scalability and system response speed.

The aim of future research is to use these methods in post-processing to filter spam and poor quality responses to develop a data set that can be used in NLP (Natural Language Processing) research.

7. CONCLUSION

A simulation of the Annotation-Validation (AV) Model shows that theoretically users can be rewarded for higher quality solutions to tasks where there is no gold standard and retrospective agreement motivates users to contribute. Analysis of data from an implementation of the model in a text annotation game indicates that, whilst useful as a way to reward users, it may not be sophisticated enough to predict user performance.

Acknowledgments

The author would like to thank the reviewers and Dr Udo Kruschwitz for their comments and suggestions. The creation of the original Phrase Detectives game was funded by EPSRC project AnaWiki, EP/F00575X/1.

8. REFERENCES

[1] M.-D. Albakour, U. Kruschwitz, J. Niu, and M. Fasli. Autoadapt at the Session Track in TREC 2010. Gaithersburg, Maryland, USA, Nov. 2010.

[2] A. Bosu, C. S. Corley, D. Heaton, D. Chatterji, J. C. Carver, and N. A. Kraft. Building reputation in stackoverflow: An empirical investigation. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 89–92, Piscataway, NJ, USA, 2013. IEEE Press.

[3] J. Chamberlain, K. Fort, U. Kruschwitz, L. Mathieu, and M. Poesio. *ACM Transactions on Interactive Intelligent Systems*, volume The People's Web Meets NLP: Collaboratively Constructed Language Resources, chapter Using Games to Create Language Resources: Successes and Limitations of the Approach. Springer, 2013.

[4] J. Chamberlain and C. O'Reilly. User performance indicators in task-based data collection systems. In *Proceedings of MindTheGap'14*, 2014.

[5] S. Deterding, D. Dixon, R. Khaled, and L. Nacke. From game design elements to gamefulness: Defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, MindTrek '11, pages 9–15, New York, NY, USA, 2011. ACM.

[6] B. Kanefsky, N. Barlow, and V. Gulick. Can distributed volunteers accomplish massive data analysis tasks? *Lunar and Planetary Science*, XXXII, 2001.

[7] H. Ma, R. Chandrasekar, C. Quirk, and A. Gupta. Improving search engines using human computation games. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 275–284, New York, NY, USA, 2009. ACM.

[8] M. Poesio, J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 2013.

[9] W. Rafelsberger and A. Scharl. Games with a purpose for social networking platforms. In *Proceedings of the 20th ACM conference on Hypertext and Hypermedia*, 2009.

[10] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, Aug. 2010.

[11] L. Schamber, M. Eisenberg, and M. S. Nilan. A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management*, 26(6):755–776, Nov. 1990.

[12] K. Seaborn, P. Pennefather, and D. Fels. Reimagining leaderboards: Towards gamifying competency models through social game mechanics. In *Proceedings of Gamification 2013: Gameful Design, Research, and Applications*, pages 107–110, Stratford, Ontario, 2013. ACM.

[13] L. von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.

[14] G. Zichermann and C. Cunningham. *Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps*. O'Reilly Media, Inc., 2011.